Sustaining Interdisciplinary Research: A Multilayer Perspective

Hultin, Alex

Award date:
2018

Awarding institution:
University of Bath

Sustaining Interdisciplinary Research: A Multilayer Perspective

Alexander Hultin

A thesis submitted for the degree of Doctor of Philosophy

University Bath

Department of Mechanical Engineering

June 2018

“Alone we can do so little; together we can do so much.”

-Helen Keller

Abstract

Interdisciplinary Research (IDR) has received a lot of attention from academics, policy-makers, and decision-makers alike. RCUK invests £3 billion in research grants each year (RCUK 2017); half of the grants are provided to investigators who hail from different departments. There is mounting awareness of the challenges facing IDR, and a large body of literature trying to establish how IDR can be analysed (Davidson 2015, Yegros-Yegros, Rafols et al. 2015). Of these, the majority have been qualitative studies and it has been noticed that there is a distinct lack of quantitative studies that can be used to identify how to enable IDR.

The literature shows that many of the barriers to IDR can be classified as either cultural or administrative (Katz and Martin 1997, Cummings and Kiesler 2005, Rafols 2007, Wagner, Roessner et al. 2011), neither of which are easily changed over a short period of time. The perspective taken in this research is that change can be affected by enabling the individuals who conduct IDR. Herein lies the main challenge; how can these future leaders of IDR be identified so that they can be properly supported.

No existing datasets were deemed suitable for the purpose, and a new dataset was created to analyse IDR. To isolate dynamics within an organisation, hard boundaries were drawn around research-organisations. The University of Bath journal co-authorship dataset 2000-2017 was determined to be suitable for this purpose.

From this dataset a co-authorship network was created. To analyse this, established models from literature were adapted and used to identify differences in disciplinary and interdisciplinary archetypes. This was done through a correlational study. No statistically significant differences between such author archetypes were found. It was therefore concluded that an alternative approach was necessary.

By adapting the networks framework to account for different types of links between edges, a multilayer perspective was adopted. This resulted in a rank-3 tensor, node-aligned framework being proposed, allowing disciplines to be represented in the network. By using this framework to construct the University of Bath multiplex co-authorship network, an exemplar structure was established through use of a series of proposed structural metrics.

A growth model was proposed and successfully recreated the structure and thereby uncovered mechanics affecting real-world multiplex networks. This highlighted the importance of node entities and the layer closeness centrality. This implies that it is very difficult to carry over benefits across disciplines, and that some disciplines are better suited to share and adapt knowledge than others. The growth model also allowed an analytical expression for the rate of change of disciplinary degree, thereby providing a model for who is most likely to enable and sustain IDR.

Acknowledgements

I would like to take this opportunity to formally thank my academic supervisors **Prof. Linda B. Newnes** and **Dr. Nigel Johnston** for their advice, constructive criticism, and encouragement throughout this research process. This research would not have been possible without their input. The support that they have given me as a doctoral candidate have helped me develop personally and professionally for which I will be eternally grateful.

I would also like to thank my examiners, **Prof. Marcelle McManus** and **Prof. Glenn Parry** for their invaluable insight and suggestions for improvement.

I would like to thank the input of **Edward Newland-Pratt** under whose guidance I learned to program in Python. He provided me with great insight into Scraping, Machine Learning, Web Development, Visualisation, and many aspects that this research required for completion. He also provided valuable verification of my implementations.

I would like to thank my wife, **Sarah Hultin**, for all the support that she provided in every possible sense of the word in addition to the valuable proof-reading and sense-checking that have turned my ramblings into something legible. She has stood by me every step of the journey.

I would like to thank my parents **Pia** and **Göran Hultin**. Not only have they supported my education and pushed me to better myself, they have also provided the loving support that only parents can give.

Finally, I would like to dedicate this thesis to those who are no longer with us, but who I know would be proud of my achievements: **Agnes** and **Sven Hultin**, **Anita** and **Erik Pohjonen**, and of course: **Tiggy**, **Teddy**, **Aydi**, and **Widget**.

List of Publications

Alexander Hultin, T. T., Nigel Johnston, Martin Kirkman (2014). Modelling Effective Product Development as Network-of-Networks. IEEE International Systems Conference, SysCon. Ottawa, Ontario, Canada, IEEE.

Alsharman, M., R. Fairchild and A. Hultin (2017). Analysing Financial Herding Through Network Analysis. 9th International Conference on Computational and Financial Econometrics. U. Senate House & Birkbeck University of London. London, UK.

Ellinas, C., M. Hall and A. Hultin (2014). DESIGN THROUGH FAILURE: A NETWORK PERSPECTIVE. DS 77: Proceedings of the DESIGN 2014 13th International Design Conference.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

ABM	Agent Based Model(ling)
AMSTAR	A MeaSurement Tool to Assess systematic Reviews
API	Application Programming Interface
CSS	Cascading Style Sheets
CSV	Comma Separated Values
DS-I	Descriptive Study One
DS-II	Descriptive Study Two
DT	Decision Tree
ERGM	Exponential Random Graph Models
HPM	Hierarchical Process Modelling
IDR	Interdisciplinary Research
KDE	Kernel Density Estimation
NLTK	Natural Language ToolKit
OLS	Ordinary Least Squares
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PS	Prescriptive Study
PSM	Problem Structuring Methods
R&D	Research and Development
RAKE	Rapid Automatic Keyword Extraction
RCUK	Research Council United Kingdom
REF	Research Excellence Framework
SCA	Strategic Choice Approach
SLRs	Systematic Literature Reviews
SNA	Social Network Analysis
SODA	Strategic Options Development and Analysis
SSM	Soft-Systems Methodology
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF	Term Frequency
TFIDF	Term Frequency, Inverse Document Frequency

Chapter 1: Introduction

Research is one of the most vital endeavours society can undertake. It is responsible for the technological advancements that we enjoy daily, our understanding of the world, and ultimately provides us with the tools to improve society as a whole, as well as being centrally important to the growth of all economies (Conway and Steward 2009, Edquist 2010, Atkinson and Ezell 2012).

The identification of Systems of Innovation has provided governments, organisations, and policy makers with a framework to discuss how research affects a country (Edquist 1997). It has been shown that research provides many benefits, including: improvements in processes and procedures (Utterback and Abernathy 1975, Kline and Rosenberg 2010), technologies that can be commercialised into products (Utterback and Abernathy 1975), development of skills and expertise needed in a rapidly developing global economy (Atkinson and Ezell 2012), and improved decision and policy making (Senge 1991).

The UK coordinates its research through its seven Research Councils under the umbrella of Research Councils UK (RCUK). It is the stated aim of RCUK to "ensure the UK remains the best place in the world to do research, innovate and grow business." (RCUK 2017). To that end, RCUK invests £3 billion per year to support research across all Research Councils to achieve this, and supports delivery of efficient, impactful research. It is estimated that research yields a return of 20-50% (Haskel and Wallis 2013), providing a net positive contribution of £600 million to £1.5 billion annually.

A major component of RCUK's strategy has been to promote interdisciplinary research across all seven research councils with the recognition that real world problems are inherently interdisciplinary, and thus require expertise to reflect this (EPSRC 2014). More than 50 percent of Research Council active grant portfolios have been reported as interdisciplinary (where the investigators come from different departments) (EPSRC 2014).

1.1. Research context – Interdisciplinary Research

Empirical studies provide support for RCUK's recognition of real world problems requiring Interdisciplinary research (IDR) (Carayol and Thi 2005, Barry, Born et al. 2008, Corsi, D'Ippoliti et al. 2010, Van Rijnsoever and Hessels 2011). It has been reported that the quantity of industrial links is greater for IDR, and that the researchers tend to have key strategic positions (Carayol and Thi 2005, Van Rijnsoever and Hessels 2011). This indicates that IDR provides direct skills to industry and provides crucial links between academia and industry (Conway and Steward 2009, Edquist 2010). RCUK is, at the time of writing, proposing a 5-year £1.5Bn fund, and is calling IDR hubs to apply for funding (Innovation 2018).

However, whilst it is recognized that IDR is important, the metrics available to determine the outcome of IDR are lacking (Van Rijnsoever and Hessels 2011, Rafols, Leydesdorff et al. 2012, Siedlok and Hibbert 2014, Davidson 2015, Yegros-Yegros, Rafols et al. 2015, Huutoniemi and Rafols 2016). There are two commonly reported problems facing IDR.

The first is a coordination problem: disciplines not sharing meanings or norms, or having organisational, cultural, and administrative differences (Katz and Martin 1997, Cummings and Kiesler 2005, Rafols 2007, Wagner, Roessner et al. 2011). This provides a significant inhibitor to effectively conduct IDR.

The second is a lack of appreciation for interdisciplinary work, as research standards are defined by individual disciplines, making it difficult to appreciate the work as it will be different from established norms from any one perspective (Phillips 2010). This raises two problems: difficulty in generalising approaches for use in future problem situations (Bruce, Lyall et al. 2004) and evaluating the success of IDR (Klein 2008). This makes it challenging to promote IDR in academia because despite its applicability, there is no definition or metric that can adequately evaluate the performance of any IDR. Furthermore, this raises issues for RCUK and its research councils. The number of citations is the primary metric that is used to measure the impact of a publication (Harzing). As IDR does not have a specific audience, the lower citation numbers have been reported as IDR not being successful (Davidson 2015). Researchers have identified this as a problem, but there have been few proposals to overcome this (Davidson 2015).

It has been suggested that an approach that measures the "integration of theory, method, or data from at least two different fields of research that reflect new insight into, or understanding of, a problem, a method, a data set, or phenomenon." should be employed (Wagner, Roessner et al. 2011). However, no such effective operational definition or measure has been found. Without a definition of what makes IDR effective, it is impossible to define which individuals are effective.

Ultimately, given that IDR is a vital component of the UK's academic and industrial organisations, it clearly has great potential. The best way to overcome its challenges and support its opportunities is to enable policy and decision makers to monitor research on an organisational level, and make evidence-based decisions. This thesis improves upon the approaches taken to assessing IDR and its key actors.

This research takes the view that effective collaborations promote repeat collaborations. Therefore, sustaining IDR is seen as a measure of effective collaborations (Mansilla, Lamont et al. 2013). In pursuit of this, this thesis develops an approach that can explore, analyse, and visualise collaborations on an organisational level, culminating in a model that identifies individuals for future IDR.

1.2. Thesis Outline

This thesis is organised as shown in Figure 1.1. This figure illustrates the relationship between each of the chapters and outlines the narrative for the overall research. Chapter 1 outlines the context of the research and introduces the problem. This provides the lens for how it is that the research should be viewed, outlines the research aim and objectives, and provides structure to thesis.

Chapter 2 outlines the research methodology. This chapter outlines the structured approach through which the research has been designed. The approach outlines the research philosophical assumptions adopted, which has great bearing on the rest of the research. A positivistic and deductive approach is taken, which specifically focuses on creating knowledge by testing and corroborating hypotheses.

Chapter 3 represents the first step in the structured research approach. It explores the context of IDR, and the various approaches taken to analysing it. It finds that there have been relatively few quantitative approaches to studying it. With unprecedented access to tools to collect and analyse big data, a Social Network Analysis (SNA) approach is deemed to be appropriate.

Chapter 4 reviews and outlines the basic theory of Networks Science that is necessary to understand SNA. It then reviews SNA studies that have been specific to IDR or analogous to IDR (e.g. research in general).

Chapter 5 outlines the requirements of a dataset and outlines that the boundaries of the research are centred on an organisation, which should be matched by the data collection process. The University of Bath co-authorship network suits these purposes well and provides a dataset that can verified. The dataset is then validated.

Chapter 6 addresses the need to identify disciplines in order to quantify whether a co-authorship is disciplinary or interdisciplinary. It brings to light that the operational definition of disciplines can be defined based on either department or the content of individuals' work. Methods are developed and implemented to extract individuals' disciplines based on these definitions.

Chapter 7 implements and analyses several models identified in Chapter 4 on the University of Bath co-authorship network that was developed in Chapter 5 and 6. The validity of these models is tested on the University of Bath co-authorship network with respect to two different metrics that indicate successful research. No statistically significant differences are found between disciplinary and interdisciplinary researcher archetypes. It is concluded that SNA lacks the resolution to suitably analyse IDR as it needs to make a distinction between different types of links: disciplinary and interdisciplinary links.

Chapter 8 reviews the nascent field of multilayer networks. It identifies two major frameworks that have been adopted in literature, identifies the challenges that the field faces, and discusses major approaches to understanding the effects that the multilayer dimension creates. The review finds that the field of multiplex network growth models can provide great benefit to understanding how it is that such multiplex networks are created. By understanding the mechanism of how multiplex networks are created, predictions can be made on which individuals are most likely to enable and sustain IDR.

Chapter 9 proposes a series of metrics that would represent various aspects of a multiplex network. These are applied to the University of Bath co-authorship network to show exemplar structures. A series of growth models are then created, and the resulting networks are then analysed with proposed metrics. Good agreement is found on very simple rules. The most striking finding is that by treating an individual as separate entities in different disciplines, good predictive capability is achieved for future IDR.

Chapter 10 discusses the analyses of the various models. The validity of the assumptions, data, and the models is discussed. Whilst the overall prognosis is positive, the application of these models need to be tempered with its limitations. It also discusses the application of the final proposed model, and outlines that it is best used to engage stakeholders in discussion and help them take evidence-based policy decisions. This has always been the strength of networks, which otherwise is best represented as a matrix.

Chapter 11 concludes the research. It proposes the research aim has been achieved but outlines the vast and important future work that still needs to be undertaken.

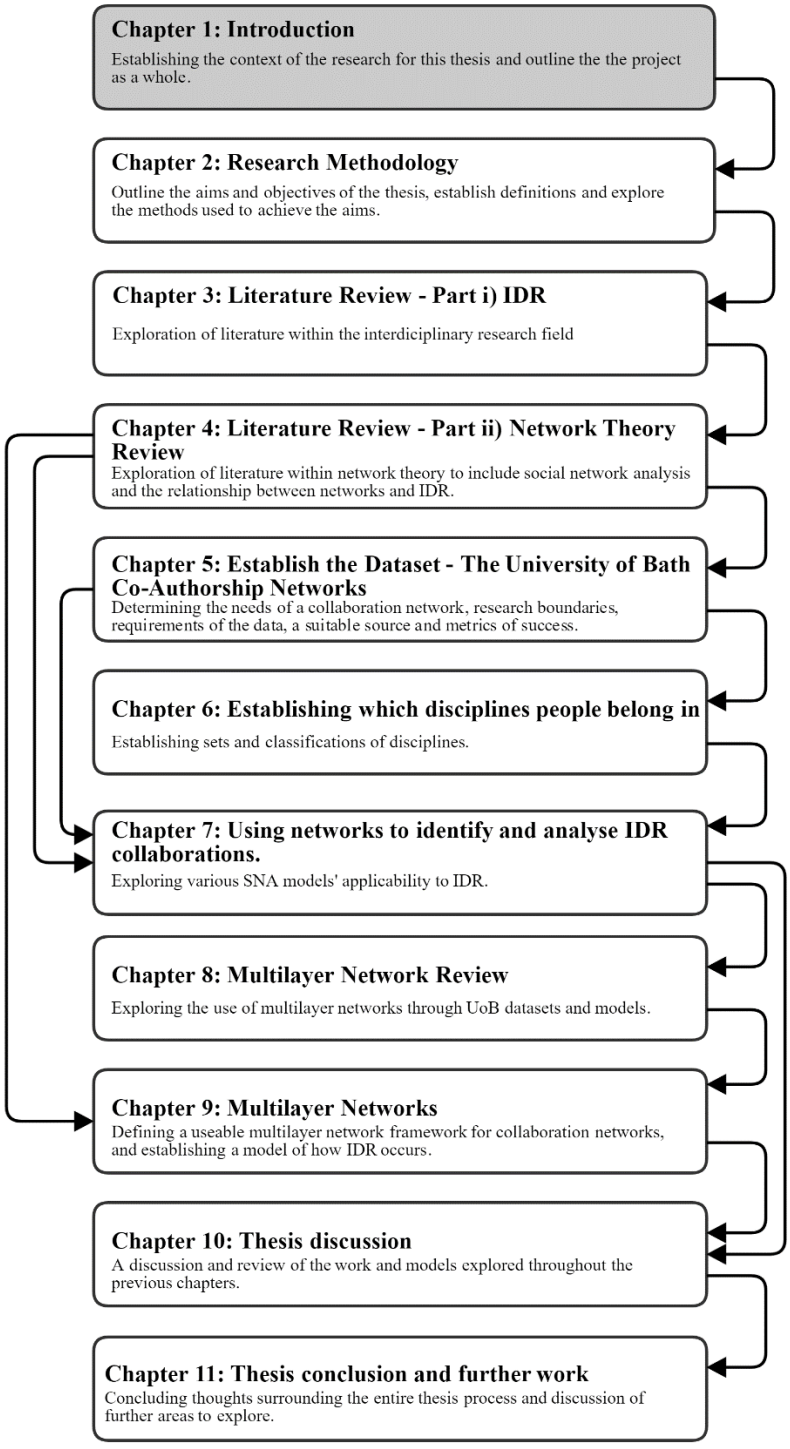

Figure 1.1: Thesis Navigation - Chapter 1 has aimed to give context to the overall thesis and provide a basis upon which the Research Methodology can be founded upon.

Chapter 2: Research Methodology

This chapter first outlines the research aim and objectives. It then discusses the needs of the methodology. The methodology framework is then outlined and discussed. This framework outlines the elements of the methodology and provides structure to how these different elements relate to each other. This chapter finally defines each of the specific elements adopted.

To continue, the research methodology needs to be defined. The definition of 'methodology' varies, but there is consensus that it is distinct from 'method'. Method is defined here as a specific procedure undertaken to answer a research question or disprove a hypothesis. Methodology has been defined as a "system of methods" (Dictionary.com 2018), "the theory of how research should be undertaken" (Saunders, Lewis et al. 2011), and "the interconnection between the applied methods" (Kreye 2011). The definition adopted here is that it is a system of methods. This includes how the methods are interlinked, and how it is that they achieve the research aim.

Whilst methods are part of the methodology, only an outline of the methods are provided here. The specific details and discussions will be provided in their associated chapters.

2.1. Aim and Objectives

IDR provides a lot of opportunities to conduct research into real world problems, approach research with different lenses, draw on knowledge from different disciplines, and increase the innovativeness of the research (Van Rijnsoever and Hessels 2011, Siedlok and Hibbert 2014).

However, it faces a lot of challenges. It must overcome cultural, administrative, and semantic barriers whilst establishing paradigms that accommodate such collaborative worldviews.

A lot of research has focused on trying to overcome these, but three major approaches are generally suggested: improved management of IDR, changes in policy to promote IDR, or a cultural change in how it is that we perceive IDR on all levels (Van Rijnsoever and Hessels 2011, Daspit, Justice et al. 2013, Yegros-Yegros, Rafols et al. 2015). The first requires a lot of resources (e.g. a research manager), whilst the other two are long-term solutions as it is difficult to reach consensus on how to achieve the desired change, and it will always take a long time for majority adoption.

These all require a significant amount of resources to accomplish and does not help us deal with making IDR effective in the immediate future.

Therefore, this research provides an alternative approach. It seeks to provide policy and decision-makers with a model that can achieve goals on a shorter term. By identifying the future leaders of IDR, decision-makers can select individuals who are most effectively able to achieve excellent IDR

outcomes. These would be individuals who have overcome the barriers that IDR poses, or are naturally predisposed to take advantage of its opportunities.

Such individuals enable IDR.

Furthermore, by identifying individuals who not only enable IDR, but sustain it too, it is possible to ensure that the research is built upon and builds lasting collaborators for future projects, further contributing to knowledge creation and reducing the barriers between disciplines.

As such, the research aim can be defined as follows.

> *To create a model that identifies individuals who enable and sustain interdisciplinary research.*

This research aim seeks to make the world a better place by identifying individuals who can effectively undertake research that addresses real world problems and grand societal problems. Furthermore, by identifying such individuals, we can enable them to further develop IDR protocols, and train the next generation of IDR researchers. This ensures effective efforts towards addressing real world problems and developing knowledge creation through cross-fertilization in both the short-term and the long-term.

To achieve this aim, several research objectives were developed and achieved. These objectives are summarised here.

Objective 1. To choose an appropriate and useful approach to analysing IDR from a people centric perspective.

Objective 2. To define and collect a dataset that suits the needs of the chosen approach.

Objective 3. To establish the validity of prevailing models in IDR literature and analogous research in the collected dataset.

Objective 4. To develop a framework that addresses the deficiencies of the prevailing models in IDR literature and analogous research.

Objective 5. To develop a model using the framework to achieve better predictive capability in identifying the future leaders of IDR in comparison to standard approaches.

Objective 6. To validate the model using the collected dataset.

Objective 7. To discuss the strengths, weaknesses, and implications of the created model.

2.2. Methodology framework

To achieve the aim a structured methodology provides means to undertake rigorous and structured research. It provides a way of choosing and defining the research design. This structured research

methodology is best defined by adapting well understood methodological frameworks. However, there is a trade-off between the structure of a research framework and its adaptability. A few different frameworks are considered. This section covers four different approaches.

1. Kumar and Phrommathed (2005) defines an eight-step linear process. This framework lays out a clear process, and is supplemented with extensive literature on how it is that the research should be conducted, and what points to consider. This highly structured approach is very clear and readable. As such, it can provide a clear description of the research process, but a complete definition of the methodology requires a well-defined research design. Furthermore, this structure is rigid, and will rarely reflect a true research path as research in real world problems is often complex (Checkland 1983).
2. Saunders, Lewis et al. (2011) do not provide a process framework, but rather provide a method defining framework named 'the research onion'. This framework provides structure on how to define the methods of the research. As with peeling an onion, each layer must be addressed before reaching the centre. Thus, the framework can be thought of as a procedure for defining the following:
    i. Research philosophy.
    ii. Research approach (deductive or inductive).
    iii. Research strategy.
    iv. Research choices.
    v. Time horizons.
    vi. Techniques and procedures.

    This provides a grounded methodology, and defines how it is that methods fit together (research choices). This framework does not provide a description of the process and is thus better suited to provide the basis of and help define the research design.

3. Blessing and Chakrabarti (2009) proposes a four-stage methodology: Criteria, Descriptive Study I (DS-I), Prescriptive Study (PS), and Descriptive Study II (DS-II). This approach seeks to understand, respectively, what the criteria of success are, what the current state of the system is, what improvements are needed based on the understanding of the state of system, and what the state of the system would be after performing the desired intervention and whether the desired outcome is achievable. This approach provides a better appreciation of how the research process occurs and how it is that the different methods interlink, but does not provide a full methodology.
4. Systems and Soft Operation Research have developed many different Problem Structuring Methods (PSMs). These have been designed to overcome problems regarding 'complex

systems'[1], or problems considered "messy" (Ackoff 1979), or "wicked" (Rittel and Webber 1973). The methodology adopted needs to be able to deal with these issues as appropriate (i.e. if the problem is simple or well defined, a simple methodology is appropriate).

There are many different PSMs that can suit the research needs (e.g. Deming cycle, Strategic Options Development and Analysis (SODA), Soft-Systems Methodology (SSM), and Strategic Choice Approach (SCA)). Ultimately, PSMs are designed to take pluralist approaches and focuses on stakeholder engagement to identify the problem, and clarify a solution space (Sterman 2000). For the purposes of defining a methodology, PSMs are very abstract concepts and make it difficult to follow.

A combined approach was determined to provide the flexibility and structure. Kumar's clear and easily relatable structure provides the structure desired (Kumar and Phrommathed 2005), this is done within the understanding that research is iterative. This iterative approach is meant to capture the cyclical nature of research, where the more we learn about a problem, the better we are able formulate and answer it (Sterman 2000). This has been captured in many PSMs (e.g. the Deming cycle as shown in Figure 2.1.

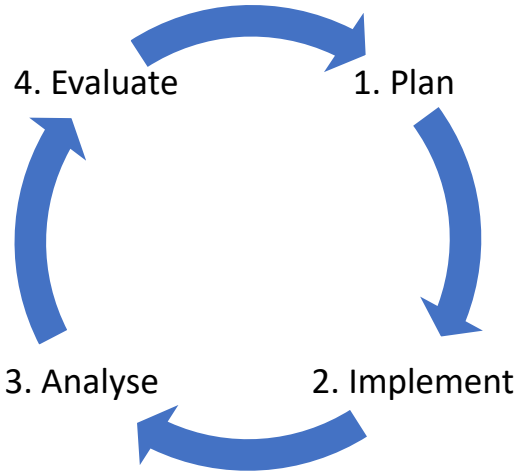

Figure 2.1. The Deming cycle consists of 4 (or 3 steps depending on the version) that seeks to establish a structured way to approach problems. It incorporates the idea that the process is cyclical as the more we learn from the process, the better we can plan how to achieve the desired results.

Furthermore, the 'research onion' is incorporated into the framework as a structured way to outline the research design (Saunders, Lewis et al. 2011). The resulting methodology framework can be seen in Figure 2.2. The framework has tried to capture the cyclical nature of research by going through studies 1 and 2 (each being represented by a single cycle of the Deming cycle).

[1]A complex system is defined as an unknowable system due to the sheer complexity of how it functions.

Step 1 in Figure 2.2, “Formulating a research aim” was done in Chapter 1. The “Literature Review” was established in Chapters 3 and 4 as precursors to the Networks study, and in Chapter 8 as a precursor to the Multilayer Network study. This chapter will continue to describe the process of “Conceptualising a research design” (Step 2). “Constructing an instrument for data collection”, “Selecting a sample”, “Writing a research proposal”, and “Collecting data” (Steps 3-6) were conducted separately, but from the same source and using improvements to the tools, and are presented jointly as their final iteration in Chapters 5 and 6. Processing the data (Step 7) for the Networks study is dealt with in Chapter 7, whilst the Multilayer Network study is dealt with in Chapter 9. “Writing a research report” (Step 8) is not discussed in this work, but the outcome is the thesis itself.

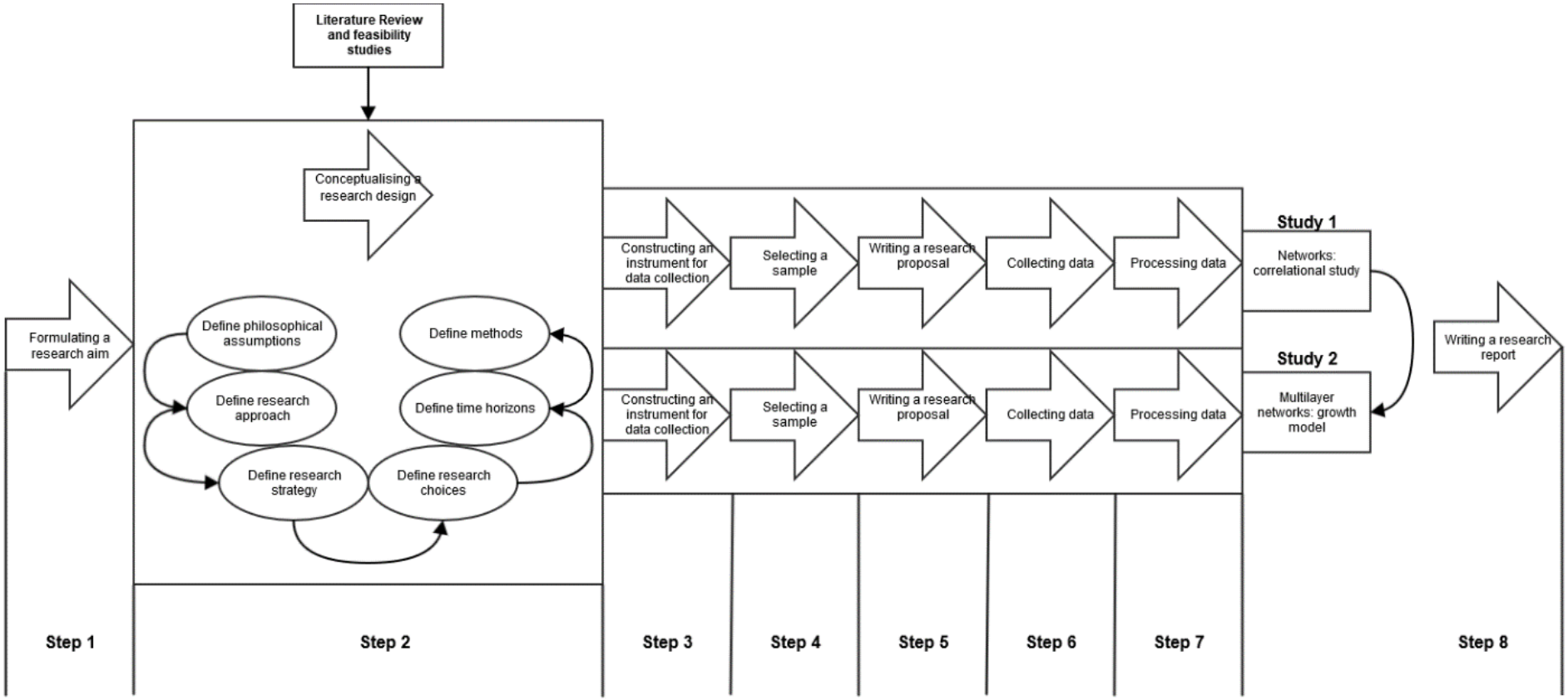

Figure 2.2. The adapted research methodology. It follows a linear process in method, but is conducted iteratively as per the Deming cycle. Therefore, the transition between Study 1 and Study 2 can be considered an evaluation step, where the researcher's worldview has been altered, and a paradigm shift requires further research. This is theoretically done until the research aim has been achieved.

2.3. Conceptualising a research design

Unlike the other steps, conceptualising a research design is not a straight forward concept. As such, it is necessary to describe what this entails, why it is important, and how it fits in with the overall research methodology.

Conceptualising a research design is defined in this research as the process of creating a research design. The output of this step should therefore be a research design.

This is important as it provides the foundation of the research. That is to say that the research design is ultimately the definition of what each study is, whereas the rest of the steps are actualising this research design.

The research onion was integrated into this step as it provides structure to this process. Therefore, the conceptualisation of the research design consists of its steps, where each step leads onto another, as if peeling an onion.

2.3.1. Research philosophy

The research philosophy outlines the basic assumptions made about the research and its environment, and serves as the first step in defining a research design. The research philosophy outlines the overall view of the world, and how it and data can be analysed. Saunders, Lewis et al. (2011) suggests that it provides a clarification of the researcher's beliefs. Outlining these can provide far greater insight into which method is appropriate.

Two research philosophy branches have been identified as being pertinent: ontology, and epistemology.

- **Ontology** is the study of the nature of reality. That is to say, does the world contain an ultimate truth? For the natural sciences, the answer is more obvious, but for questions dealing with social constructs, it is less obvious. For instance, "iron has a melting point of 1,538°C" is true regardless of whether humans observe it or not. However, "Steve is a good manager" is not necessarily true for everyone.
- **Epistemology** is the study of knowledge and perhaps one of the more famous branches of philosophy. The tripartite argument (knowledge is justified, true belief) is one of the most commonly used definitions of knowledge. The 'research epistemology' outlines what constitutes acceptable knowledge. Such knowledge will come in the form of being able to express or evaluate a theory, or theoretical proposition.

Four philosophical positions have been identified to provide a position on these two branches (Saunders, Lewis et al. 2011):

- Positivism
    - A philosophy with the basic affirmation that knowledge is based on "positive" data of experience, logic, and mathematics. It is therefore strictly worldly, antitheological, and antimetaphysical. A number of different iterations of positivism exists (e.g. logical positivism, critical positivism). For the purposes of this research, a generic definition of the philosophy assumes that the universe exists independent of the research, that only "positive" data of experience, logic, and mathematics provide credible data and facts, and that knowledge can be affirmed through these.

      The positivist epistemological position has gained a lot traction due to the strength of Falsificationism,, whose most famous proponent is Karl Popper (Thornton 2017). This is due to positivist condition that knowledge needs to be testable, and experienced. It is for this reason that hypothesis testing is usually associated with a positivist epistemology.
- Realism
    - A philosophy that generally assumes that reality exists independently of the research, but can be interpreted through social conditioning. Data and facts as with positivism, can be achieved through observable phenomena, logic, and mathematics. It is worth noting that Realism has different schools of thought that disagree on important aspects – e.g. Direct Realism attributes inaccuracies in sensation to insufficient data, whereas Critical Realism views that sensations are open to misinterpretation.
- Interpretivism
    - A philosophy centred on distinguishing between the natural realm and human realm. By virtue of the human realm being perceived, interpreted, and described, an empathetic perspective is needed on different levels of aggregation. It understands that observed phenomena are embedded in the axioms and paradigms of the social actors. It has been described as ant-positivism as it attempts to move away from a tendency to try to generalise phenomena to law, which can often be ill-suited to deal with the complexities of social systems.
- Pragmatism
    - A philosophy that argues that no one philosophical assumption is well-suited to all research questions. In research methods, it is therefore suggested that every research questions adopts a philosophy that is best-suited to answer that specific question.

The research philosophy plays an important role in determining the validity of research and therefore plays an important role in the research design. The research philosophies all have merit and therefore adopting a research philosophy is about choosing one that best suits the research needs.

The needs of this research are defined by the research aim. The research aim sets out to create a model to identify individuals who enable and sustain IDR. The various components provide different requirements on the research:

- "To create a model that identifies individuals…"
    - Repeatability of the results (i.e. those identified should not change).
    - A worldly dataset with clearly defined and homogeneous features to achieve repeatability.
    - A reduction of the features to a generalised law that identifies the future leaders of IDR.
- "… individuals who enable and sustain interdisciplinary research"
    - The existence of individuals who objectively enable and sustain IDR, i.e. the existence of objective future leaders of IDR.

Based on these premises, a positivist philosophy is well-suited to this research, and is therefore adopted.

2.3.2. Research Approach

Having defined the working research philosophy, it is possible to discuss the research approach. The research approach determines how the research is conducted procedurally. Kumar and Phrommathed (2005) and Saunders, Lewis et al. (2011) both suggest that this manifests in a choice of whether an inductive or deductive approach is taken. This section will first establish advantages and disadvantages of inductive and deductive approaches. It then establishes the deductive approach as the working research approach, but understands that inductive approaches have their uses in forming theoretical propositions, even though they may not be considered a scientific finding.

The deductive and inductive approaches can be thought of as processes. The deductive research approach seeks to present laws to explain and predict phenomena (Collis and Hussey 2013). The process starts with a theory and finishes with a hypothesis test or examining the results (Robson 2002). The inductive approach starts with an observation and finishes with a theory (Easterby-Smith, Thorpe et al. 2008, Saunders, Lewis et al. 2011, Bryman and Bell 2015). The difference is shown visually in Figure 2.3.

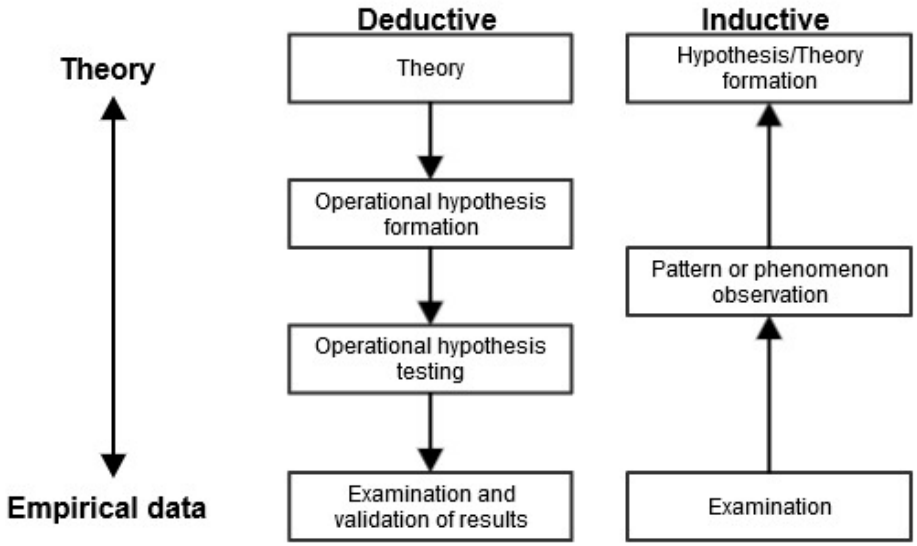

Figure 2.3. Deductive and Inductive approaches in research. Adapted from (Robson 2002) and (Easterby-Smith, Thorpe et al. 2008) respectively.

Deductivism is usually associated with a positivist philosophy, with Falsificationism being a central principle of the approach (Saunders, Lewis et al. 2011). Falsificationism is the concept that a scientific statement needs to be falsifiable. It is through this process of falsifying scientific statements, and the amendments to prevailing theories that theories become 'less bad' (Andersen and Hepburn), which has been argued as being the key mechanism to developing scientific knowledge throughout history, as proposed by Kuhn's scientific paradigm (Lakatos 1978, Kuhn 2012). To paraphrase Einstein: "No amount of experimentation can ever prove me right; a single experiment can prove me wrong" (a paraphrase of a translation of A. Einstein (Einstein 1918–1921)). The difficulty of Deductivism does not lie in the validity of the knowledge that it produces (the resulting knowledge is robust), but rather in how hypotheses are formed in the first place.

It is for this reason that the 'scientific method' is normally described as hypotheses testing of a theoretical proposition. Therein lies the central tenet of Deductivism: a hypothesis is tested, and it either holds, is refuted, or the test was statistically insignificant.

However, Deductivism has been criticised as not being able to represent complex systems as it does not methodologically provide the opportunity to understand a system, merely test it. The view taken in this work is that it can represent complex systems, but can only create knowledge about its inputs and outputs. If these are designed to understand the system itself, there is no reason Deductivism could not be and has been used to represent systems (Maurer 2007). Furthermore, the criticism bases itself on the fundamentally different purposes, inductivism seeks to understand a system, Deductivism seeks to test it.

The strength of this argument has culminated in modern philosophers coming to terms of non-provability (Greenland 1998).

Inductivism is a broad epistemological subject that has been debated at great length by the likes of Hume, Kuhn, Carnap and Popper (Hume 2003, Kuhn 2012, Carnap 2014, Thornton 2017). It remains a greatly divisive subject. Inductivism is usually defined as "the philosophy of drawing a generalizable law from observation", although it is more correct to call this *enumerative induction* or *universal inference* (Henderson).

The inductive approach has been criticized, most notably by Popper (Thornton 2017) going so far as to claim that inductive reasoning does not exist and that it is logically invalid. This stems from "Problem of Induction". Briefly described: for an inductive statement to be true, there is a necessity for the past to predict the future (e.g. as Bertrand Russel illustrates: a turkey is fed every day. Then according to Inductivism, the turkey will be fed the following day. This holds true until the turkey is slaughtered for Thanksgiving).

However, it has been argued that induction does not need to assume this repeatability and should not seek to predict unobserved instances (Greenland 1998). To gain knowledge on the unobserved requires a formulation of a hypothesis that predicts the unobserved, and is then tested under Popper's falsifiability paradigm. Induction is thus proposed as a hypothesis (or theory) formation mechanism, but with no predictive or generalisation capability.

This is the interpretation of many contemporary researchers and is considered a strength (Easterby-Smith, Thorpe et al. 2008, Saunders, Lewis et al. 2011, Bryman and Bell 2015). It is noted that inductive research is beneficial by being able to account for unforeseeable phenomena, does not simplify complex situations to simple cause-effect links to certain variables, and allows an in-depth understanding of nuance situations to be created. However, in the Popperian view, this is not considered scientific as ultimately, it has not been subjected to falsification, and cannot have the scientific rigour that is needed.

Additionally, complete rejection of inductivism can create serious difficulties. For instance, it would have led to the falsification of the conservation of energy in the 1920s during beta decay experimentation. Instead, Inductivism was used to propose the existence of the neutrino (which was only was detected in 1956) (Maher 2010).

Popper, of course, did not naively think that hypotheses could not be formed, but rather held that there is a distinction between scientific knowledge and ordinary knowledge (Thornton 2017).

The approach adopted in this research is Deductivism. This is chosen for two main reasons. First, the rigour surrounding Falsificationism provides a clear validation approach (i.e. either a hypothesis is corroborated for the dataset or it is not). This makes it easy to discuss and review. Second, the

research aim seeks to repeatably identify future leaders of IDR using a set of features. This can be thought of as generalising a law concerning IDR. Such a law is best tested using Falsificationism.

This then provides a robust approach, and a way of establishing the premises to the research by codifying them as hypotheses.

However, it is important to note that Inductivism can be useful as a way of reflecting upon findings.

2.3.3. Research strategy, choices, and time-horizons

Having established the research approach desired (deductive with some inductive elements). It is possible to examine possible research strategies, choices, and time horizons according to which are the most appropriate for the chosen research approach and that most readily achieves the research aim.

Firstly, three different types of studies have been identified across Saunders, Lewis et al. (2011) and Jackson (2014). These are exploratory, descriptive, and explanatory. Kumar and Phrommathed (2005) adds correlational studies to this list. Finally, modelling can be done for a variety of reasons, yet consists of its own challenges methods, and is therefore considered separately here (Epstein 2008).

- **Exploratory studies** seek to clarify, assess, and gain new understanding of phenomena (Saunders, Lewis et al. 2011). These are described as preliminary in nature or used to develop or refine tools or procedures (Kumar and Phrommathed 2005); exploratory studies being the main purpose of the research are less common.
- **Descriptive studies**' purpose is to provide an accurate description of a situation (Robson 2002). This may precede or follow an exploratory study (Saunders, Lewis et al. 2011), but will have to be followed up with an explanatory study to make the research meaningful (Saunders, Lewis et al. 2011).
- **Correlational studies** seek to discover the relationship between two or more variables (Kumar and Phrommathed 2005). This includes drawing inferences from confounding variables (Didelez 2007). Jackson (2014) makes a distinction between predictive and explanatory studies, stating that predictive studies seek to use correlational and quasi-experimental methods, but one cannot ascribe cause and effect explanations. This is subsumed into the correlational study paradigm in this thesis.
- **Explanatory studies** seek to establish and explain the causal relationships between variables (Saunders, Lewis et al. 2011). As causality is difficult to establish, explanatory studies tend to use experimentation as a research method. Causality is very difficult to

establish in other methods, although not impossible (e.g. longitudinal studies or other time-series datasets) (Didelez 2007).

- **Modelling studies** have traditionally been conducted to provide predictive capabilities to real world problems. However, it can be used for many other purposes, such as illuminating the core dynamics of a system (Epstein 2008) or even drawing causality about real phenomena (Larsen, Thomas et al. 2014).

The deductive research approach makes correlational studies (via statistically significant hypothesis tests) and explanatory studies (via experimentation or simulation) the most viable options. Furthermore, the research aim requires some predictive capability.

An explanatory study to ground the predictive capability may seem necessary, but an experimentation study is exceedingly difficult to conduct in realistic settings. This is true as the motivations in controlled experiments change and are usually small scale (in comparison to an entire research organisation).

For this reason, the research design is focused on investigating data that represents IDR.

Correlational studies using temporal aspects have been used to establish causality (particularly in finance) (Zaremba and Aste 2014). This can however be quite a tenuous approach. It is more usual to define that correlation studies (with or without a temporal aspect) describe the inputs and outputs. It can therefore be thought of as a "black-box" representation of the system (i.e. it describes the input and output, but not the mechanism behind it).

For the purposes of developing a model, it may be able to identify who enables and sustains IDR, but not why. This is not deemed to be rigorous enough.

A mixed-method approach was determined to be appropriate in bridging the deficiencies that pure correlational studies have. A modelling study could help understand the underlying reasons as to why some phenomena are seen (Larsen, Thomas et al. 2014).

As such, a longitudinal population study is required to establish predictive correlations. Fortunately, with increased computational capabilities, and the availability of rich data, a broad historical longitudinal study is easily achievable today (provided that the variables of interest are available). This means that correlational study is deemed to be appropriate. This is complemented by a modelling study that seeks to provide insight into why the correlations are occurring.

Two studies were conducted: one which established whether previous SNA models hold and the second a multilayer framework to address flaws in the first study. Whilst evaluating the previous SNA models, three of the five tested models held, but the approach chosen suffered from a flaw

that provides some explanation as to why few SNA IDR studies exist. As no models are directly transferable, a modelling study is designed to recreate the observed phenomena. The model is correlated to the longitudinal data and finds excellent predictive correlations.

At each of these steps, hypotheses are tested as a way identifying and codifying scientific knowledge. The hypothesis tests are tested on statistically significant trends, tested to a 0.05 significance level.

2.4. Research design

The research design therefore adopts a positivistic philosophy that is best represented by a deductive research approach. Hypothesis testing provides a clear way to communicate what is corroborated knowledge and what remains conjecture and observation. A correlational study performed on a historical longitudinal dataset can establish predictive capability and could be used to draw explanatory insight by virtue of its temporal component. This is complemented by a simulation model, validated over the longitudinal data to provide the final model and insight.

2.5. Chapter summary

This chapter has outlined the aims and objectives of the overall research. It has stated the importance of outlining a structured research methodology to not only guide the research, but to provide clarity to the readers.

An iterative methodology framework was defined, highlighting an inability for a highly structured approach to capture the research process. The iterative methodology was able to show the research process leading onto a complementary study. This can be thought of as outlining the explicit research process as the result of a Deming cycle.

As outlined in the second step of the framework, a research design was defined, which was achieved by adopting a research onion approach.

Chapter 3: Literature Review – Part i) Interdisciplinary Research

This chapter outlines the literature surrounding the challenges and opportunities facing IDR. By analysing these, it is possible to develop a lens for how to achieve the overall research aim. The aim of this chapter is thus to establish the possible approaches to achieve the research aim. To that end, a literature review has been conducted.

However, IDR has been studied with views from many different disciplines. Therefore, it is necessary that the depth and breadth of these disciplines are properly represented in the literature review. To achieve this in a structured manner, a Structured Literature Review (SLR) is created.

The SLR in this chapter seeks to achieve the following review objectives.

1. Identify through the literature the definitions of IDR and select a definition for use in this research.
2. Identify the challenges and opportunities facing IDR.
3. Identify various approaches taken to enable and overcome the barriers to IDR.

Figure 3.1 demonstrates the hierarchical relationship of objectives and concepts explored in this chapter. The top of the hierarchy provides the definition of IDR, which guides the understanding throughout the research. The ‘Opportunities and Problems’ branch provides a lens through which we can understand the costs and benefits of conducting IDR. This provides context to any research regarding IDR. The ‘Approaches to investigate IDR’ branch uses the concepts developed in challenges and opportunities to propose the various approaches that are possible to instigate change in IDR. Finally, it was identified that quantitative approaches to investigating IDR are lacking, and proposes SNA as being a viable framework to investigate a research organisation. This final branch is dealt with in Chapter 4.

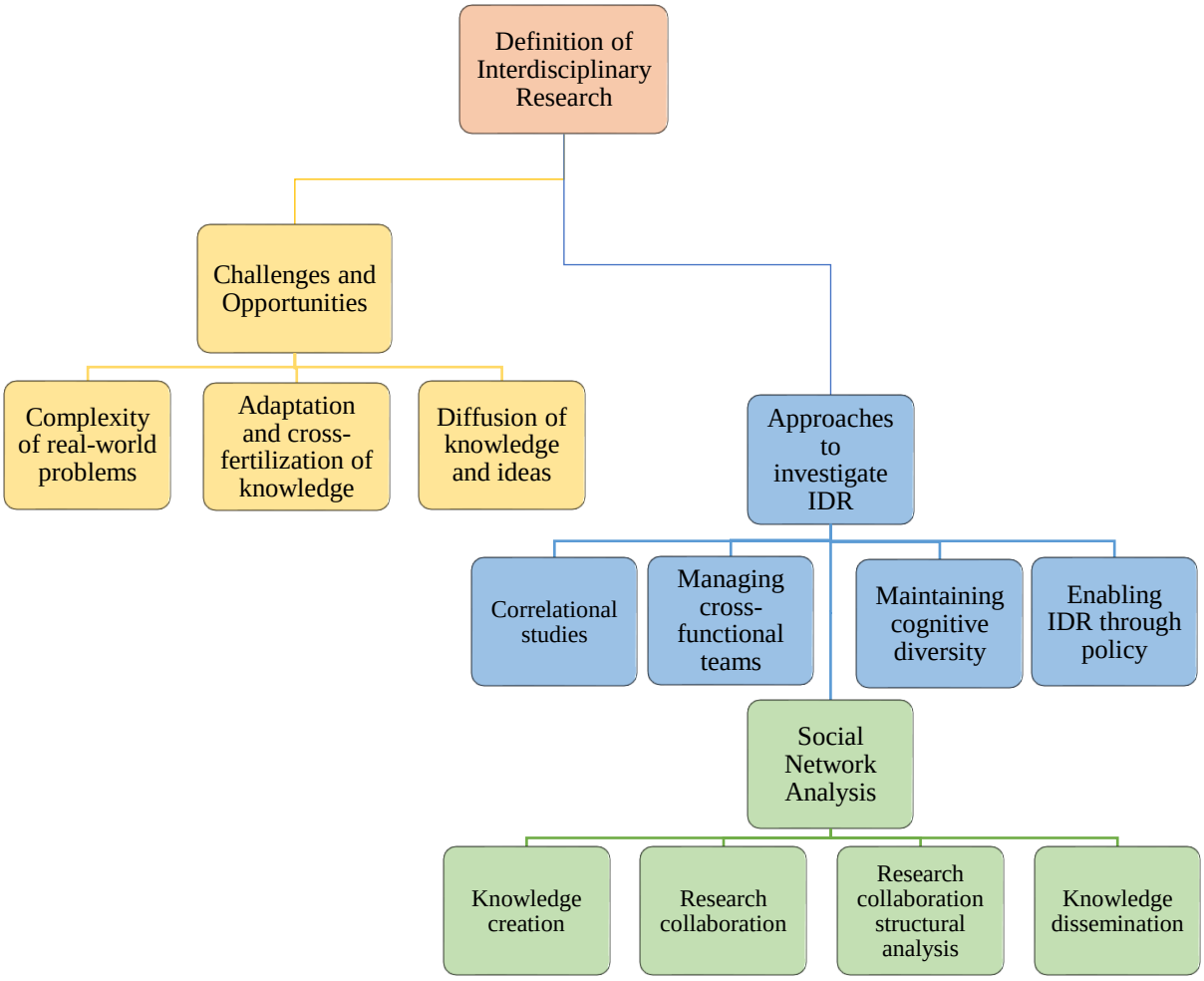

Figure 3.1. A hierarchical view of the adopted review structure in this chapter. Yellow elements represent contextual reviews. Blue elements represent reviews on approaches taken. Green elements represent analogous Social Network Analysis reviews.

Section 3.1 (‘Literature review approach’) of this chapter defines the SLR. The subsequent sections are organised according to the hierarchy shown in Figure 3.1. Section 3.2 (‘Interdisciplinary Research: Definitions’) outlines the adopted definitions of disciplinary collaborations. Section 3.3 (‘Challenges and opportunities of IDR’) provides a contextual lens to reviewing IDR literature. It specifically provides costs and benefits which should be borne in mind when reviewing approaches. Section 3.4(‘Approaches to improving IDR’) provides an overview of the major approaches taken to reviewing IDR or analogous collaborations.

3.1. Literature review approach

This section describes the approach taken to reviewing the literature. This is important it is necessary to adequately capture the breadth and depth of the fields which relate to this research.

In this research a structured approach was used to assist in meeting the required breadth and depth. Broad subjects such as IDR cover many fields, hence many factors affecting IDR will be contained within the broader field of 'collaboration', it becomes more necessary to ensure the breadth of the subject is captured.

'Systematic Literature Reviews' (SLRs) have provided approaches to answer specific research questions to ensure the repeatability of literature reviews. However, in research where exploration of a concept is necessary, such methods are not entirely suitable.

An alternative structure is proposed in this research. First, it accepts that exploration in research will alter the mental models of the world. As such, the presented literature review adopted an iterative process. For this, the Deming cycle (Plan-Do-Check-Act) provided a useful framework to address such iterations.

Second, the iterations were done in batches of papers. That is to say, search terms of interest were established, papers were then identified through a given portal, the papers were then screened, the remaining papers were then codified to establish the definitions, approaches, claims, conclusions, and findings (although other notes were made about the papers where needed), the papers were then evaluated with respect to how they affected the overall research and how they altered the researcher's mental model of the world. This in turn gave rise to new search terms, which in turn restarted the cycle.

At times, where a new concept was being explored, the snowball sampling method was useful in finding seminal papers.

The formal methodology of reviewing papers in this research consisted of the following activities conducted cyclically:

- **Scoping** is defined as choosing the research question(s) that have not been answered before. The approach taken in this research proposed that research boundaries are considered as an alternative when no explicit questions can be posed.
- **Planning** is the process of conceptualising search terms from the defined questions/boundaries. The search terms should consider the research aim, key concepts, measures/variables, research design, participants, and time frames. Equally, the paper inclusion and exclusion criteria should be defined.
- **Identification** is the systematic approach to find these papers.

- **Screening** is the process of identifying which works from those identified are appropriate for the research.
- **Eligibility** is simply defined as the process of extracting the appropriate information and determining whether the publication passes the inclusion/exclusion criteria.

  Evaluating is defined as the act of reviewing the findings. It should be noted that this requires collating and drawing concepts and findings from the papers together, and not simply summarising the papers. This process should be clear in whether it needs to be quantitative or qualitative. Quantitative papers mostly apply to finding specific statistics and will be prescriptive in its nature. Qualitative seeks to explore and develop theories, methodologies, approaches, and metrics.

The resulting method is visualised in Figure 3.2.

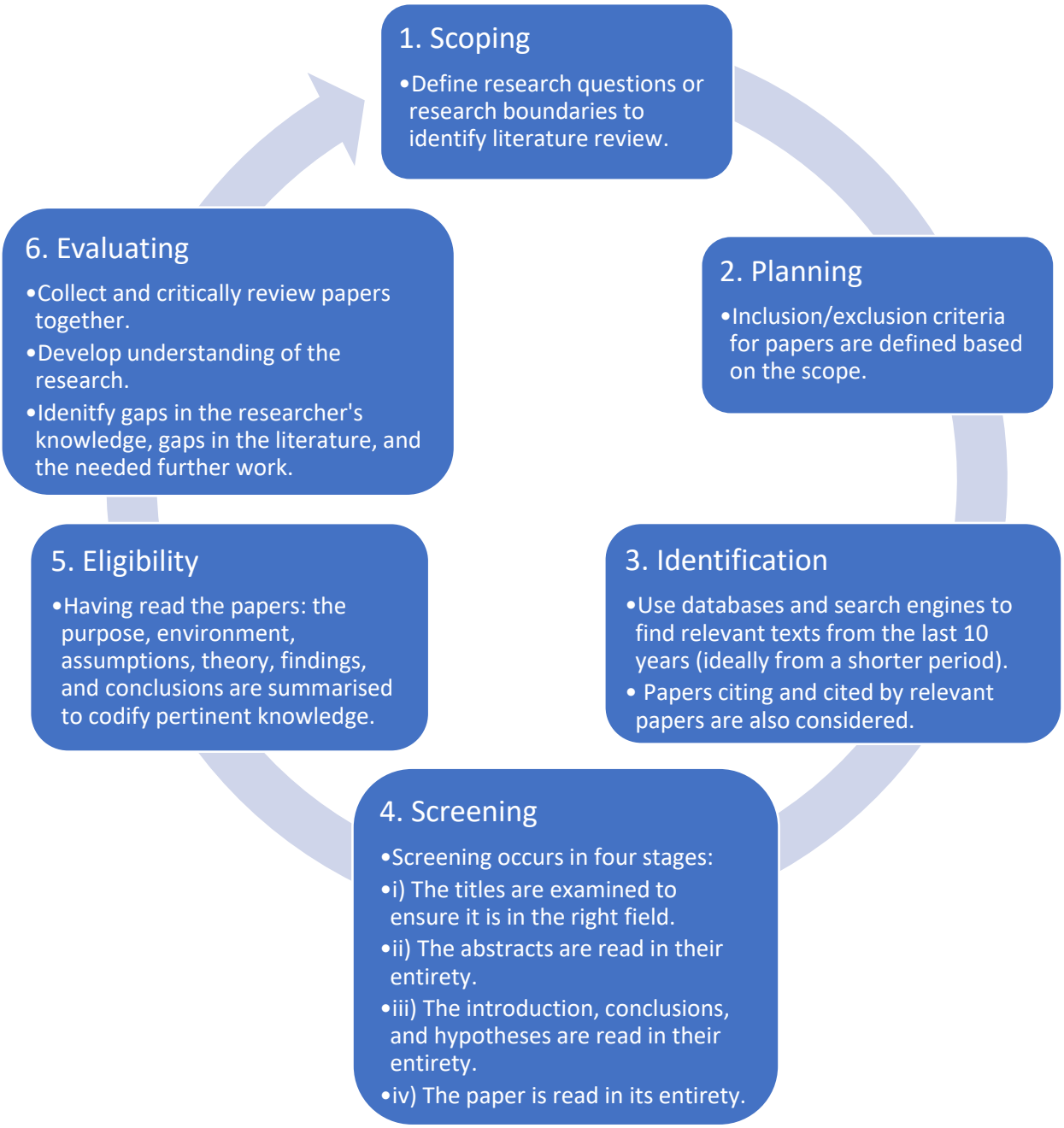

Figure 3.2. The adopted method represents an iterative cycle that is necessary to incorporate the changes in the researcher's mental models of the world. The initial starting point is the research aim, which is determined at the start of the research. The cycle should finish after step 6, although the number of cycles is uncertain and is usually determined by a combination of whether the research aim has been achieved and time constraints.

This provides the overall structure as to how the literature review was conducted throughout the research.

3.2. Interdisciplinary Research: Definitions

The first part of the literature review provides a definition of what IDR is. Oxford dictionaries defines discipline as “A branch of knowledge, typically one studied in higher education”(2017). Whilst disciplines as constructs are well understood, there is no standard operational definition. That is to say that the taxonomy of knowledge is not standardised and will vary from organisation to organisation. This becomes more complicated as Mechanical Engineering is a discipline, yet within Mechanical Engineering there are many different “disciplines” (e.g. vibrations, hydraulics, structural analysis). This complexity is further compounded by there being significant overlap between disciplines (e.g. both Physics and Mechanical Engineering contain Fluid Dynamics studies; to which does it belong or is it a discipline by itself?).

The operational definition dealt with in Chapter 6, wherein it is proposed that disciplines can be defined by either department-based disciplines, or by content-based disciplines.

The term interdisciplinary would therefore suggest that it is ‘between disciplines’, but the overall literature tends to use it interchangeably with crossdisciplinarity, and multidisciplinarity. Strictly speaking, when distinguishing between the different types of collaborations that can occur, the following definitions are used in this research.

- **Intradisciplinary** and **Disciplinary** research (used interchangeably) is defined as being research conducted within a single discipline (Cohen and Lloyd 2014).
- **Multidisciplinary** research is defined as research being conducted by two or more individuals from different fields. The work is divided into clear sub-parts, such that the disciplines do not overlap (Zeigler 1990).
- **Crossdisciplinary** research is defined as research being conducted across fields (e.g. a physicist conducting research in biology) (Zeigler 1990).
- **Interdisciplinary** research is defined as research being conducted by two or more individuals from different fields. Unlike multidisciplinary research, the work being conducted uses the ideas from all fields to achieve the research aim. The synergy of the approaches produces research that is greater than the sum of its parts (Zeigler 1990).
- **Transdisciplinary** research is research that is conducted to create a unified theoretical framework. This can be thought of as creating a new field (e.g. systems engineering) (Leavy 2011).

However, this assumes that the outcome of collaborations can be classified into one of these. This would be extremely difficult to achieve and is outside the scope of this research. This research provided greater benefit when defining IDR as the union between Multidisciplinary, Crossdisciplinary, Interdisciplinary, and Transdisciplinary definitions.

3.3. Challenges and opportunities of IDR

Having established what IDR is, it is necessary to identify the challenges and opportunities of IDR, which provides a lens through which we can understand how to properly enable and sustain IDR. This review also takes advantage of studies focusing on analogous concepts.

3.3.1. Opportunities

Sir Francis Bacon is attributed with stating "scientia potential est", "knowledge is power" (Bacon 1864), in his Meditationes Sacrae (1597). Knowledge has been defined as being central to economic development (DEVELOPMENT 1996), with patent growth being causally linked to the growth of the GDP in G7 economies (Josheski and Koteski 2011).

IDR is widely believed to provide many benefits to the development of knowledge, and ample literature can be found on the subject (Yegros-Yegros, Rafols et al. 2015). Despite the strong push for IDR, it is a difficult concept to measure, and has even been reported recently as to lacking concrete evidence of benefits (Jacobs and Frickel 2009).

Three main arguments for IDR have been identified in literature:

1. It has been argued that IDR provides improved ability to tackle new and complex problems (Davidson 2015).
   With the exponential development of science and technology since the second half of the twentieth century, few technologies today can be fully understood by any one person (Conway and Steward 2009, Kossiakoff, Sweet et al. 2011). This makes IDR a vital part of our society. Furthermore, IDR has been identified both by policy makers (RCUK 2017), and by academia (Rittel and Webber 1973, Ackoff 1979) as being necessary to tackle many real-world problems. It is for these reasons that there has been a surge of interest in IDR (Yegros-Yegros, Rafols et al. 2015).
2. One of the most common concepts across all fields dealing with integrating different knowledge is that of cross-fertilization (Conway and Steward 2009).
   This is the concept that knowledge created draws benefit from the synergy of the multiple knowledge bases. For example, signals processing, control theory, telecommunications, and mathematics all draw knowledge from each other to create a synergy (Ogata 2002).
3. IDR provides desirable and relevant skills for the public and private sector.
   National Systems of Innovation (NSIs) and the associated Triple Helix model shows us that there is a close link between universities and the private sector. This relationship is centred on the private sector benefitting from knowledge and skills developed in academia (Lundvall and Johnson 1994, Edquist 1997, Lundvall 1998, Lundvall, Johnson et al. 2002,

Lundvall 2007, Etzkowitz 2008, Edquist 2010). If IDR is used to answer real-world problems, then it stands to reason that IDR skills are desirable.

This section outlines the most relevant literature to these three aspects.

3.3.1.1. The complexity of nature and society

The complexity of nature and society means that no one field provides sufficient knowledge to fully understand all aspects (National Academies of Sciences and Medicine 2005, Manring 2014). This surprisingly becomes more evident in the natural sciences where the fields of physics, chemistry, and biology are exhibiting more and more overlap (e.g. molecular biochemistry). In engineering, the increasing complexity of new technologies has necessitated greater integration amongst knowledge bases (making systems engineering necessary (Kossiakoff, Sweet et al. 2011)). Furthermore, it has been argued that system-wide performances cannot be measured in isolation, (Blockley and Godfrey 2000, Davidson 2015). Therefore, any improvements on an overall system requires a systemic approach. As it has been argued that most real-world problems can be considered complex, the importance of including multiple worldviews and to conduct IDR becomes apparent (Marcella, Carlo et al. 2010, Phillips 2010, Willmott 2011).

Studies have found that cognitive diversity provides improvements in numerous different settings (Page 2007). A diversity of perspective, interpretation, and models provide better tools to approach complex problems. Furthermore, diversity provides a mitigation to ignorance or being unable to approach a problem by hedging knowledge (Stirling 1998, Stirling 2007).

As such, it is recognised that real-world problems need an interdisciplinary approach (Merz, Friedrich et al. 2006). The recognition of IDR being necessary to solve real-world problems coincides with the RCUK's position (RCUK 2017).

3.3.1.2. Drawing benefits from different knowledge bases

Innovation literature identifies that cross-functional teams and drawing inspiration from other knowledge bases in order to address a problem can yield highly effective solutions. Innovating firms such as the Edison Labs (Conway and Steward 2009) and IDEO (Hargadon and Sutton 1997) have advocated that horizontal communications between individuals, teams, and functions have been the basis of their innovative success (Steve Conway and Steward 2009). Interdisciplinary fields are generally accepted as providing solutions with a high novelty (Dogan and Pahre 1990, Bartunek 2007). Functional diversity has also been shown to provide improvement in performance, development times, new venture performance, and schedule performance (Eisenhardt and Tabrizi 1995, Simons, Pelled et al. 1999, Horwitz 2005, Li and Zhang 2007).

As society is complex, it has been argued that responsible innovation requires engagement with its stakeholders, resulting inevitably in a need for IDR (Taebi, Correlje et al. 2014).

In academia, IDR is also seen as a source of innovativeness as it introduces alternative paradigms to established fields. The recombination of knowledge from fields allows established knowledge to evolve into new and purposeful knowledge. (Molas-Gallart and Salter 2002, Leydesdorff, Nightingale et al. 2011). This is achieved through various mechanisms. Access to different expertise, instruments, methods, and heuristics, and promoting greater productivity, and cross-fertilisation across disciplines, and pooling knowledge have all been found when investigating the success of IDR (Katz and Martin 1997, Melin 2000, Molas-Gallart and Salter 2002, Bozeman and Corley 2004, Rafols, Leydesdorff et al. 2012). The different knowledge bases also allow for greater modularisation of tasks (Raasch, Lee et al. 2013).

Many scholars have found that the evolution of research fields is partially dependent on creating, recombining, and reutilising knowledge from different fields (McCain 1998, Tsai and Wu 2010).

3.3.1.3. Greater diffusion of knowledge between the private sector and research institutions

IDR has been well established as being vital to answering real-world problems and engineering challenges. It is important to also mention the benefit that closer integration between academia and the private sector can have. The various representations of the UK according to its triple helix model and its NSI places the Western European countries, including the UK academic system as the main provider of highly skilled labour (Etzkowitz and Ranga 2015), source of innovation (Ito, Kaneta et al. 2015), and basic research. Additionally, IDR has been shown to draw greater experience from outside academia (Van Rijnsoever and Hessels 2011).

The relationship is synergistic. The private sector stands to gain access to cutting-edge research, researchers from whom they can learn from, and highly-skilled individuals whom can provide an effective approach to a company's problems (Conway and Steward 2009). The entrepreneurial professor is a phenomenon that can provide huge benefits to the private sector, both in terms of creating products and services (Conway and Steward 2009) and in establishing hubs of innovation such as the Silicon Valley (Conway and Steward 2009). The development of core scientific knowledge enables future IDR to establish uses for such technology - e.g. graphene (Zurutuza and Marinelli 2014).

Academia, on the other hand, benefits from greater exposure to the private sector. Such exposure ensures that academia aligns itself well with the needs of the private sector. This provides two major benefits: it ensures that the research is relevant, and it provides insight into the skills that are needed so that the next generation of students can be taught sought after skills (Edquist 2010). Finally, the

private sector is a rich source of data and resources that are vital to academia (Schwartz and Vilquin 2003).

3.3.2. Inhibitors to IDR and their associated costs

Whilst most of the literature is overwhelmingly in favour of conducting IDR, it is generally agreed upon that there are several issues that it needs to overcome in order to gain greater traction and mainstream appeal.

These inhibitors cause real costs that must be borne by the researchers and their funding bodies. These costs come in the form of inefficiency in conducting IDR as teams and individuals. Team assembly and coordination is the subject of many studies (Cummings and Kiesler 2005, Rafols 2007). Such coordination is especially difficult in teams with diverse backgrounds (Michalisin, Karau et al. 2007, Wagner, Roessner et al. 2011). For instance, the definition of successful IDR has been reported to be different from member to member of the same IDR research teams (Roche and Rickard 2017). Disciplines often do not share meanings and norms (organisational, cultural, and administrative), making the barriers to successful IDR difficult in many cases (Davidson 2015). As such, the transaction costs between IDR members is higher than intradisciplinary teams (Nooteboom 2000).

Furthermore, IDR is difficult to conduct with a lack of appreciation, research standards, and an audience (Rafols 2007, Rafols and Meyer 2007, Barry, Born et al. 2008, Van Rijnsoever and Hessels 2011, Yegros-Yegros, Rafols et al. 2015). The structure of research organisations tends to be split into departments based on disciplines. The function of such departments is to establish standards and metrics (Yegros-Yegros, Rafols et al. 2015). This includes output metrics that are used to determine the success of researchers. This becomes especially problematic for IDR in the UK as research success metrics are based on quality journal publications (e.g. the UK has adopted the 'Research Excellence Framework' (REF), which seeks to classify submissions into a star-based rating system (Martin and Whitley 2010)). The use of such a rating system nationally opens up the avenues for comparisons between individuals and organisations (Taylor 2011), and foster unhealthy practices, such as only hiring individuals who subscribe to a popular ideology, limiting discourse (Rafols, Leydesdorff et al. 2012).

Bibliographic measures are defined as measures of the quality and quantity of research outputs in the form of peer-reviewed publications. Most approaches have been citation based. Articles are ranked based on the journals they are published in by the impact factor of the journal (Bornmann and Daniel 2008), or the preferred journals as outlined by departments (Martin and Whitley 2010). Authors are usually measured by several different indices based on the citations of their

publications. The H-index is one of the more commonly used indices, which H publications with at least H citations each (Hirsch 2007). However, such indices have several weaknesses.

One of the most limiting aspects of the H-index is that citations grow over time, and so will the H-index without there necessarily being any more papers published. An M-index has been proposed to counter this, which is simply the H-index divided by the number of years since the first papers was published (Harzing , von Bohlen und Halbach 2011). The H-index can also ignore very highly cited papers, which the G-index tries to overcome by finding the largest value of G having at least G^2 citations (Egghe 2006). Equally, the E-index tries to overcome the ignored citations by doing the opposite; finding the E publications with at least $\sqrt{E}$ citations.

There are many variations, but they all suffer from the same criticisms.

The H-index is difficult to compare across fields. Even within a discipline, it is often not a fair comparison as certain aspects of a discipline may have a wider audience than others (Bornmann and Daniel 2008, Anauati, Galiani et al. 2016). The H-index also provides integer values, resulting in a loss of data resolution (Ruane and Tol 2008). This is further skewed by the fact that the H-index is affected by self-citation (Bartneck and Kokkelmans 2011, Emilio and E. 2013). There is also some question as to how to detect all citations as no database is complete. Google citations crawls through scientific documents and records these automatically, whereas other databases require the authors to be recognised. Both are made difficult where the author publishes under different format of their name (e.g. J. Smith, John Smith, or Smith, J. A.).

Despite these biases, the use of the H-index and other similar indices is wide-spread as a performance metric and used for policy decisions.

Within each of these arguments, there is a bias against fields that are less cited. Without a stable audience, IDR journals are not highly cited (Klein 2008), nor can they usually produce generalisable findings as they are often based on complex real-world problems (Bruce, Lyall et al. 2004). These disparities make it difficult to evaluate the quality of IDR from an outside perspective. This has resulted in IDR being less rewarding than disciplinary research in award recognition, and career advancement (De Boer, De Gier et al. 2006, Levitt and Thelwall 2008, Siedlok and Hibbert 2014). Despite there being significant interest in IDR (Melin 2000, Whitley 2000), there is a concerning tendency for IDR researchers to revert to discipline-based research (Raasch, Lee et al. 2013).

It is worth noting that certain other measures have reported better predictive capabilities. Something as simple as the mean number of citations per author has better indication of future performance (Lehmann, Jackson et al. 2006). Other authors question the purpose of the H-index as they found a strong correlation with the H-index and the number of papers published ($H \sim 0.54\sqrt{n}$) (Yong 2014). Adopting the number of papers would certainly circumvent the citation problem in bibliographic

measures. Other studies have used simpler measures to obtain, such as the sum of authors' publications' impact factors (McFadyen and Cannella 2004).

One of the more famous ranking systems is a researcher's Erdős number, which is the distance in citations of a researcher to world-renowned mathematician Paul Erdős. It has been proposed that the structure of the citations should also be taken into consideration giving rise to several robust bibliometric measures. One of the most successful rankings is the Phys Author Rank Algorithm (Dixit, Kameshwaran et al. 2009), which uses an eigenvector centrality-based approach to map the ranking of scientist across time. Therefore, it is possible to create bibliographic measures that could provide fairer representation for IDR.

3.4. Approaches to improving IDR

Having established the challenges and opportunities of IDR, it is now possible to review the major approaches to taken to instigating change by overcoming the barriers of IDR, or taking advantage of the opportunities.

It is important to note that the review highlights the various approaches that have been taken, which have been primarily focused on cultural and policy changes, which require a significant amount of time to implement, require a lot of buy-in from major stakeholders, and requires a unified (or at least agreed upon) approach (Van Rijnsoever and Hessels 2011). Other approaches propose better management, which require management resources (especially in skill) (Daspit, Justice et al. 2013). As such, the ethos taken in this research is one of minimal resource usage. Therefore, the literature regarding approaches to IDR is reviewed with the difficulty of implementation in mind and is one of the reasons that an interpretivist approach was not adopted.

3.4.1. Managing cross-functional teams

Studies focusing on cross-functional teams have a wide range of findings. Cross-functional, horizontal links have been reported as being vital for innovative capability in seminal works (Freeman 2013), and have started a host of research into cross-functional team and inter-team dynamics (Conway and Steward 2009), where organisations must overcome the "creative tension paradox" (providing structure and freedom for creativity) (Peters and Waterman Jr , Zaltman and Duncan 1977, Lawrence and Lorsch 1986).

Many studies have investigated this, with recommendations having a lot of commonality with identifying the inhibitors. For instance, Hollaender, Loibl et al. (2008) suggests that effective transdisciplinary research can only be achieved by facilitating mutual learning, creating synergy

through integration of interest and goals, and stimulating mutual adjustment. Bruce, Lyall et al. (2004) investigates the EU Fifth Framework Programme (FP5), a programme aimed to increase integrated research, and found that team-building and giving time to develop appropriate semantics is vital for effective IDR to occur (instead of the multidisciplinary research that was predominantly found).

Other approaches have sought a more holistic approach to address the deficiencies in educating people to work in inter or transdisciplinary teams (Klein, Grignon et al. 2004), or through addressing the issue of IDR and transdisciplinary research not being as well established by proposing greater engagement with journal editors and creating a college of transdisciplinary researchers (Kueffer and Hadorn 2008). Whilst these may improve the cross-functional integration, these are very long-term projects.

The field of management provides a host of studies that investigate cross-functional team dynamics. Many other positive influences have been reported, for instance: greater team and inter-team cohesion has shown to improve their respective effectiveness (Daspit, Justice et al. 2013), provide overall improved performance (Keller 2001), provide faster product development times (Eisenhardt and Tabrizi 1995), and has a greater propensity to create effective solutions (Hargadon and Sutton 1997), whereas some studies have found a negative influence (Ancona and Caldwell 1992, Simons, Pelled et al. 1999, Bunderson and Sutcliffe 2002). More recently there has been a focus on the effectiveness of cross-functional teams to remain agile and how this can be achieved. Cross-functional teams have performed better at new venture performance (Li and Zhang 2007), where small cross-functional teams form a vital component of agile strategies (Rigby, Sutherland et al. 2016, Abrahamsson, Salo et al. 2017).

Adaptive management practices have been proposed for IDR to better manage teams. (König, Diehl et al. 2013) proposes to repurpose the Competing Values Framework (Quinn 1988) for IDR to integrate established management practices with the case dependent needs of IDR teams. It is with frequent team changes in IDR, it has been shown to be necessary to repeatedly re-invent the wheel (König, Diehl et al. 2013).

These works have found that effective integration and management of diverse teams provides greater ability to remain responsive to the market, organisational, design, and product needs (Parker 2003).

3.4.2. Benefitting from cognitive diversity

Other approaches have taken a knowledge, and cognitive resource diversity approach, where cognitive diversity fosters new knowledge. This has its basis in the work of Thomas Kuhn's *The*

Structure of Scientific Revolutions, which argues that scientific knowledge grows on established theories, and disproving an established theory causes a period of scientific upheaval as alternative theories are formed (Kuhn 2012). IDR exhibits the features associated with cognitive diversity and its benefits (Horwitz 2005, Barry, Born et al. 2008, Conway and Steward 2009).

Granovetter (1973) establishes the 'strength of weak ties' where he argues that the fewer avenues of interaction two people have, the less likely it is that these people have similar knowledge and can thus learn from each other. This is based on the theory of homophily, and has been extended to many different fields, such as knowledge management, innovation research, sociology, and psychology (Granovetter 1973). It is argued that collaboration between individuals is best for innovative outputs. This is supported by other works that provide both a theoretical basis (Page 2007), as well as empirical evidence (Yegros-Yegros, Rafols et al. 2015) as to why 'diversity trumps ability'.

Soft operations research advocates the use of pluralist approaches to approach complex problems too, as complex problems require multiple perspectives to be effectively approached (Vennix 1999, Mingers 2011). PSMs provide such approaches. Mingers (Mingers 2011) argues that: PSM is a rigorous and structured approach, it allows a consolidation of worldviews without falling into single measure, encourages stakeholder engagement, uncertainty is expected and tolerated and that they (the PSMs) *"aim for exploration, learning and commitment rather than optimization"*. The author reviews the different PSMs available. SSM seeks to unify worldviews to reach consensus by engaging with a conceptual model (Checkland 1999). Several models have sought to achieve pluralist perspectives through engagement and group model building. System dynamics has been used to approach complex problems as well as unifying approach (Sterman 2000). Furthermore, it uses as a group model building exercise to achieve consensus is well documented (Rodrigues and Bowers 1996, Andersen, Richardson et al. 1997, Vennix 1999, Rouwette, Vennix et al. 2002, Eskinasi, Rouwette et al. 2009). Hierarchical Process Modelling (HPM) has also been established as an approach that can represent a whole system and engage pluralist perspectives (Mujtaba 1994, Davis, MacDonald et al. 2010).

3.4.3. Enabling IDR through policy

A large body of literature has approached the issue of IDR from a policy perspective, and how bibliographic measures cause IDR to be at a disadvantage (Van Rijnsoever and Hessels 2011, Rafols, Leydesdorff et al. 2012, Siedlok and Hibbert 2014, Davidson 2015).

Despite the significant amount of literature on the subject, few quantitative studies have been published (Van Rijnsoever and Hessels 2011, Rafols, Leydesdorff et al. 2012, Raasch, Lee et al.

2013, Van Noorden 2015, Yegros-Yegros, Rafols et al. 2015). The studies that have been performed are all in agreement and show that journal ratings perform significantly better for disciplinary research.

3.5. Proposed approach: Identifying the collaborations of least resistance

From the reviewed research, it can be seen that many of the approaches specific to IDR are long-term goals, or require careful management and highly skilled managers, which may not be feasible in an academic setting. Ultimately, these relate to how it is that people collaborate, and therefore the approach should be collaboration-centric.

Many different approaches could be taken to research this issue. However, the research aim and the philosophy adopted require a quantitative approach that can be developed into a model that can be used to identify individuals in the future.

This requirement narrows down the choices as this requires a data source that is easily collected, and that can be developed into a model.

SNA has provided a framework for many different studies of varying motivations, aims, and outcomes (Wasserman and Faust 1994, Newman 2010, Barabási and Pósfai 2016). It provides a definitive lens to quantitatively study any system where a relationship exists between two objects.

Seeing as IDR seeks to understand how individuals collaborate across disciplines, this approach seems tailor-made for the research aim. As such, networks are further explored in Chapter 4.

3.6. Summary

This chapter has outlined the literature review method adopted throughout this research. The review method allows the breadth and depth of a topic to be analysed, understood, and allows for changes in worldviews.

This chapter first reviews the opportunities that it presents and the challenges that IDR poses. Many studies have found that creativity and innovativeness increase with the cross-fertilization of knowledge. Furthermore, applying approaches from different fields can yield unique perspectives, whilst applying different paradigms can help advance science. IDR has also been reported as being vital to addressing real-world problems and therefore holds a vital position in the relationship between academia and industry. However, conducting IDR is hampered by administrative, cultural, and semantic barriers. Equally, sometimes meshing paradigms from different fields can be difficult. Furthermore, as there is no stable audience for IDR, IDR can be difficult to generalise and be made

useful to academics (are papers about specific IDR collaboration applicable to all IDR?). Works related to IDR therefore have been reported to receiving fewer citations, the main metric on which academic performance is judged.

The approaches to improving IDR are based on managing and enabling people's abilities. It is for this reason that a collaboration-centric perspective is adopted in this research. As such, SNA is deemed to be a suitable and effective approach.

However, traditional sociological methods make it difficult to establish quantitative approaches to identifying these individuals, making it an expensive approach that requires bespoke investigations (Hamill 2006). Few quantitative approaches exist, and fewer still with readily available data. With ever increasing digitalisation of the academic environment and the development of analytical tools, it becomes ever more feasible to take advantage of big data analytics.

These studies show that studying research organisations is possible using networks as a representation of these people's abilities, prominence, research interest, and propensity to collaborate (Wasserman and Faust 1994). This chapter seeks to establish whether IDR, propensity to conduct IDR, and the state of IDR can be established through the use of networks.

As such, a black-box view is taken in identifying these individuals. That is to say that it is very difficult to be able to identify all the factors that contribute to an individual being able to facilitate IDR. It is outside the scope of most organisations capabilities to do this continuously and on a large scale. A more resource effective approach is to identify statistical properties that are strongly correlated. Further work could then include understanding why these properties are correlated, and disseminating the specific factors that drive these properties (e.g. what factors affect a person's degree centrality?).

The literature clearly shows that network statistical properties in collaboration networks represent a wide number of factors including ability, sociability, prominence, and organisation embeddedness. Furthermore, with the advent of big data availability, the cost of producing such networks is cheap, and can be easily be updated to have a "live feed" of the research organisation's collaboration network.

Chapter 4: Literature Review – Part ii) Network Theory Review

Chapter 3 established that due to the major challenges and opportunities lying within how it is that individuals collaborate across disciplines, SNA is a natural approach to achieve the research aim.

This chapter reviews the Social Network Analysis approaches that have been taken to study research systems. Research systems are studied as there is a lack of SNA studies focusing on IDR specifically. To review this literature, it is first necessary to review network notation, structure, and statistical mechanics as these need to be understood to appreciate the SNA literature.

4.1. Origins and notation

Networks science has had a resurgence of interest in the last two decades with ever increasing computational capability, and unprecedented availability of data. However, the field itself has existed for a long time in the form of graph theory. A network is a set of common 'nodes' interconnected by a set of common 'links', whereas a graph is a common set of 'vertices' interconnected by a set of common 'edges'. The two differ in purpose and semantics, but the basis and mathematical operations are identical.

It is important that nodes and links are commons sets. For instance, a network can consist of people connected by their relationships to one and other, but that same network cannot contain a computer (unless serious advancements are done in the field of artificial intelligence and computers can be classified as people!). This simplicity can be used to represent many different types of systems with studies ranging from the internet, worldwide web, friendship networks, and email connectivity to protein chains, Bose-Einstein condensates, and food chains (Albert and Barabási 2002, Newman 2010, Barabási and Pósfai 2016). The information that can be drawn from such networks range from information about any one node, or the pattern of these networks. Both provide us a wealth of knowledge regarding the nature of that system, and direct applicability (e.g. finding the most connected person at an organisation). The strength of networks is that it provides a measurable topology that enables mathematical procedures to be applied. This includes calculating statistical properties of a network (Albert and Barabási 2002), as well as simulating its dynamics (Barzel and Barabasi 2013).

The earliest known paper on graph theory was authored by Euler (1741). The paper showed that it was impossible to cross each of the seven bridges in Koningsberg without crossing one of the bridges at least twice (Euler 1741). Graph theory started in earnest with Erdős and Rényi (1959) analysis of random graphs. A graph was defined as N vertices with E edges connecting the nodes, defining a graph $G(N, E)$. This can be written as a matrix, A, with dimensions $N \times N$, and the elements of $A_{i \in N, j \in N}$ represent the edges, E, between vertex i and neighbour j. Most studies tend

use binary values for the $A_{i,j} \in \{0,1\}$; 1 if a relationship exists, 0 if it does not. Most studies tend to be undirected as well, where $A_{i,j} = A_{j,i}$. In such a graph, there are $N(N-1)/2$ unique possible edges. It is possible for the element values to be weighted, which can represent the intensity of the relationship.

This forms the basis still used today (although weighted networks utilise $A_{i,j} \in \mathbb{R}_{>0}$, and directed networks do not require $A_{i,j} = A_{j,i}$). This can then be manipulated to provide topological measures such as centralities (how central a node is), clustering (how clustered a node or network is), and the significance of shortest path lengths.

4.1.1. Centrality

Centrality is a core concept in networks science. It provides a topological measure of how central a given node is to the network itself. This can be thought of as which nodes are the most important to a network. The distribution of the centrality measures can provide great insight into the overall topology, both on its own and in comparison (Newman 2010). Centrality provides valuable information about any network, but for research networks, it can provide valuable information on how influential or important an academic is to the overall network.

Degree centrality is the most commonly used centrality measure. (Erdős and Rényi 1959) defined the degree of a node, *k*, as the number of unique edges that belonged to a node. This provides an easily calculable and effective measure (Freeman 1977, Freeman 1978). Degrees have since been used as a powerful proxy to represent the structure of graphs. Degree distributions have been established as way of analysing the structure of graph, and forms the basis of network-wide analysis (Bollobás 1981). For randomly connected graphs, the degree distribution has been shown both analytically and stochastically to form a Poisson distribution. Such a distribution can be analysed and compared on several well-established properties such as peak height, standard deviation, skewness, and tail properties (Newman 2001, Albert and Barabási 2002). However, more recent findings have found that real networks do not exhibit such a distribution and instead follow a power-law where the exponent is the major property (Barabási and Albert 1999, Albert and Barabási 2002, Newman and Girvan 2004, Barabási and Pósfai 2016). However, whilst this centrality provides a lot of information about the structure of the network, it provides no further indication other than the number neighbours a node has. As a centrality it can be quite limiting by not providing any information on how close it is to other nodes.

Other measures overcome this by determining how well a single node reaches all other nodes. The **Closeness** Centrality is such a measure. It measures how close a node is to all other nodes in the network, and is calculated as equation 4.1 (Bavelas 1950).

$$C_{closeness_i} = \frac{1}{\sum_j shortest\ path\ distance(j)} \tag{4.1}$$

Betweenness Centrality is a similar measure, and uses the number of shortest paths that go through the node (Freeman 1977).

$$C_{betweeness_i} = \sum_{j_1 \neq j_2 \neq i \in G}^{N} \frac{number\ of\ shortest\ paths\ going\ through\ i}{number\ of\ shortest\ paths} \tag{4.2}$$

Both centralities can be thought of as indicators of how well information propagates to/from the specific node. The difficulty with using these centralities is that it requires N^2 shortest paths to be found. These are incredibly computationally expensive, and can be difficult to implement, however they have been extensively used to analyse academic collaboration networks (Brandes 2001).

Eigenvector Centrality is given in equation (4.3) (where n is the largest positive eigenvalue of $\boldsymbol{A}$ that satisfies the Perron-Frobenius theorem). The centrality is arguably the most versatile and effective measure in terms of calculating how central a node is (Barabási and Pósfai 2016), as it accounts not just how central a given node is in its immediate vicinity, but also how central its neighbours are (which in turn calculate its neighbours) (Keener 1993). The Perron-Frobenius theorem ensures that there is a unique, positive eigenvector. The centrality has been used in many different applications such as ranking College American Football teams (Keener 1993), and ranking webpage results like Google. Various different implementations of the eigenvector centrality exists, such as the Google PageRank (Page, Brin et al. 1999), and Katz (Katz and Martin 1997) approaches. The PageRank approach is particularly robust and does not require conditions *ii* and *iii* of the Perron-Frobenius equation.

$$C_{eigenvector_i} = \frac{1}{\lambda_n} \sum_{j}^{N} \boldsymbol{A}_{ij} \cdot C_{eigenvector_j} \tag{4.3}$$

The Perron-Frobenius Theorem states:

> *If M is an n x n nonnegative primitive matrix, then there is a largest eigenvalue λ_0 such that*
>
> *i.* *λ_0 is positive.*
> *ii.* *λ_0 has a unique (up to a constant) Eigenvector $\boldsymbol{v_1}$, which may be taken to have all positive entries.*
> *iii.* *λ_0 is non-degenerate*

iv. $\lambda_0 > |\lambda|$ *for any eigenvalue* $\lambda \neq \lambda_0$.

4.1.2. Clustering

Clustering is a concept that seeks to identify groups of highly interconnected nodes, or how dense a network, or sub-graph is. This is important to IDR because it provides a measure of how unifying an individual is, how closely connected a network is, or how tightly knit communities are.

The first attempt at determining this was undertaken in Erdős and Rényi (1959). This paper used a clustering expression for the entire graph given in (4.4).

$$C = \sum_{i}^{N} \frac{2 \cdot k_i}{N(N-1)} \tag{4.4}$$

The analysis found that at certain key probabilities (depending on the graph size), this value increased very rapidly. However, this has been criticised as being better described as a measure of density (Newman 2010). Watts and Strogatz (1998) adopt a similar approach, but improve on the measure by calculating the clustering on an individual basis (4.5) and then finding the average for the entire network.

$$C_i = \frac{2 \cdot k_i}{k_i(k_i - 1)} \tag{4.5}$$

Watts and Strogatz (1998) showed that randomly rewiring a regular ring lattice created local clustering, triadic closures and hubs, which also significantly reduced average path lengths, and was able to replicate the so-called "small-world" network property. The average clustering value was used to demonstrate that real networks have a higher clustering value than randomly connected networks. This is a result of random networks not generating local clustering, triadic closures, or a realistic degree distribution (Watts and Strogatz 1998). However, it is important to note that the degree distribution produced still creates a Poisson-like distribution, and the algorithm is therefore not representative of real networks.

It could be argued that a better measure of clustering should consider the clustering structure. A measure that finds the proportion of triangles in connected triplets provides a simple measure (Wasserman and Faust 1994).

$$C = 3 \cdot \frac{number\ of\ closed\ triplets}{number\ of\ connected\ triplets} \tag{4.6}$$

Equation (4.6), suggests measuring it as the ratio of triangles (closed triplets) to the number of connected triplets with only 2 edges between them.

4.1.3. Network path-lengths

Paths in networks have been used in many different contexts and play a central role in network science (Barabási and Pósfai 2016). A path is the connections taken to reach from one node to another. The shortest path is the most direct path (i.e. with fewest node jumps between two nodes), and its length (the number of jumps) is called the distance, although it is usually just named path-length (as the distance is usually the only length that is of interest). The longest distance in a network is called the diameter of the network and provides information on how closely connected a network is. Average path-lengths provide a similar measure to the diameter, but due to its non-integer nature, provides a greater resolution. This is particularly important if a structural change occurs in a network; such approaches have been taken to investigate cascading failures e.g. Italy's blackout in 2003 (Parandehgheibi and Modiano 2013, Ellinas, Hall et al. 2014).

4.1.4. Topology of real networks

Centrality, clustering, and path lengths have provided networks science measures that can be used analyse the overall structure of networks. Small-world networks are also characterised by a much smaller average path-length and diameter (Albert and Barabási 2002). However, as stated above, the distribution of centralities has provided networks science with the ability to analyse and compare network structures. A rewired lattice network still exhibited Poisson like degree-distributions (Albert and Barabási 2002), like random networks. The resurgence of network science occurred with the seminal discovery of true network topology following a power-law distribution and not a Poisson distribution (Barabási and Albert 1999). (Barabási and Albert 1999) found that measured network topologies do not exhibit Poisson distributed degree distributions, but rather they follow a power law approximated by equation (4.7), where γ is between *2* and *3*.

$$P(k) \sim k^{-\gamma} \tag{4.7}$$

This relationship has been named the scale-free property, as this relationship does not change with the scale of the network (unlike randomly connected networks). It has also been defined as heterogeneous networks, as there is large difference in degrees, whereas random networks are called homogenous for the opposite reason. Simply put, in real networks very few nodes are very well connected, whilst the majority are poorly connected. This has been confirmed in the world-wide web, the internet, email networks, cellular networks, protein chains, and sexual contacts networks

(Albert and Barabási 2002). This sparked a range of new theories on how networks are formed (prevailing theories centre on preferential attachment, wherein a preferred node is more likely to attract new connections, usually on the basis of the number of previous connections – "the rich getting richer" effect (Albert and Barabási 2002)), as well as explaining certain previously observed phenomena such as the small-world properties. It also has serious implications on failure of networks, where a targeted failure on the most connected node will have a very sudden and very fast failure i.e. breakdown of the network structure (Ellinas, Hall et al. 2014).

4.1.5. Network matrix analyses

Whilst these properties provide the basis of many cross-sectional studies, matrix analyses provide a great deal of interesting properties.

As was noted above, the eigenvector corresponding to the largest eigenvalue of a primitive matrix provides a highly useful centrality measure. However, the other eigenvectors provide information about the structure of the network as well. The second largest Eigenvalue is commonly known as the algebraic connectivity. This is a measure of how difficult it is to partition a network. Its corresponding Eigenvector, the Fiedler Vector, bisects the network into two partitions based on the sign. Further partitions can be achieved by partitioning the sub-graphs (treated as their own networks). This is highly useful when investigating IDR, as it provides a means of identifying clusters within and between departments (if they exist). This is a simple method for automatic community detection, but provides no measure of the best number of communities. As such, community detection has been of interest, particularly in social sciences where categorising individuals as belonging to a certain clique is an immensely powerful tool

Algorithms have been developed to automatically detect the number of partitions and partitioning for a network, which achieves a partition identical to the analytical approach, but are computationally very expensive ($O(mN^2)$ and $O(N^2)$) (Newman 2003, Newman and Girvan 2004, Newman 2006). More modern techniques such as the Louvain algorithm achieves similar high accuracy at a fraction of the cost (Blondel, Guillaume et al. 2008).

The graph spectral density is calculated using a network's N eigenvalues, λ_j, of the adjacency matrix, A. The spectral density (density of states) can then be calculated as $p(\lambda) = \frac{1}{N}\sum_{j=1}^{N}\delta(\lambda - \lambda_j)$. This is important as it is directly related to the topology, with the kth moment giving the number of paths leading to a node. This means that the 2nd moment will give the number of edges, $\left(E(G) = \frac{1}{2}\sum\lambda_j^2\right)$, whilst the 3rd moments will give the number of triangles $(T(G) = \sum\lambda_n^3)$, etc… (Butler 2008). This is particularly useful when calculating the clustering of a network (how closely-knit communities are), or other path dependent measures. For smaller path

lengths, this can equivalently be done by multiplying the matrix by itself, (A^x), the resulting matrix gives the number paths of length *x* that exist between *i* and *j*.

The spectra of graphs can also be used to provide important information. It can also be used to determine the structure of the overall network. For instance, it can be used to determine the number of unique spanning trees.

4.1.6. Weighted networks

The field of networks has benefitted from its simplicity. It provides a robustness to the findings, which can easily be replicated. However, such simplicity results in the loss of potential information as it does not discriminate between different relationships (as only binary networks have been considered thus far).

The inclusions of weight to networks has provided a quantitative measure to determining the strength, intensity, distance, or frequency of the relationship between the two nodes. Weighted networks are of particular interest in social interactions, where relationships are not equal. A simple example would be frequency of communication between a close friend and an acquaintance. By not considering the weighting in such cases, a significant amount of information and analysis resolution is lost.

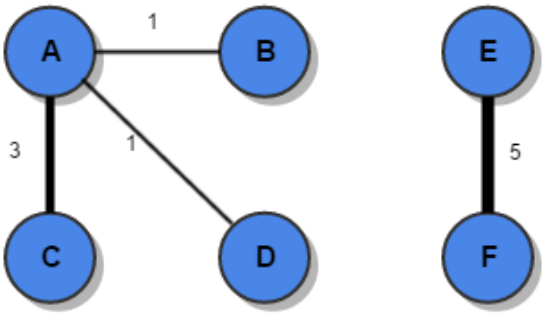

Figure 4.1. Examples of weighted networks. Node A and nodes E and F have the same strength centrality if the weighting is not normalised. To not be able to distinguish between the relationships between A and C from B and D would lessen our understanding of the network.

It is important to note that using distance as the weight is best treated as an inverse weight, which has its applicability in 'travelling salesman' optimisation problems (e.g. routes (Junjie and Dingwei 2006)). Weighted networks do not differ significantly from unweighted networks. It is denoted by

the matrix, W, where the elements are given by $A_{i\in N, j\in N} \cdot w_{i\in N, j\in N}$, where w is the weight of the connection (which could be thought of as just *1* in unweighted graphs) (Opsahl 2009). This does not change the way many things are calculated, but it can significantly change the findings.

For instance, Barrat, Barthélemy et al. (2004a) and Barrat, Barthelemy et al. (2004b) investigated the structure of weighted networks using strength (the weighted degree, s) distribution (analogous to the degree distribution). It was found that the total strength of a node in the networks followed a fat tailed distribution. Furthermore, it is correlated to the degree as per equation (4.8).

$$s(k) \sim k^{\beta} \tag{4.8}$$

If $\beta = 1$, and the distribution is well correlated, the strength of a node is dependent on the degree and is simply multiplied by the average weight of the network, $s(k) = k\langle w_{ij}\rangle$. Alternatively, if $\beta > 1$, the weighting plays a role in the structure of the network. Specifically, the strength of the node develops faster than its degree.

It was discovered that Scientific Collaboration Network was correlated very well by $\beta = 1$, suggesting that the weight of the network does not significantly alter its structure. Conversely, the Worldwide Airport Network was better fitted by $\beta = 1.5 \pm 0.1$, implying that the weights develop faster than the degree of node.

The weight also provides data that can provide greater insight into the structure of the network. For instance, a distribution was found that the edges belonging to nodes *i* and *j*, $\langle w_{ij}\rangle$, can be approximated by $\langle w_{ij}\rangle \sim \left(k_i k_j\right)^{\theta}$. For the Scientific Collaboration Network, there was no major change, with lower average weights for higher degree pairs. However, for the Worldwide Airport Network, the exponent was positive showing that hubs attract greater weight.

Other approaches have attempted to determine the importance of weight using the strength, although none have provided a measure as scalable as the above (Barthélemy, Gondran et al. 2002, Menichetti, Remondini et al. 2014).

It has been argued that the weights and strengths of links and nodes respectively should be weighted appropriately, where $0 \leq \alpha \leq 1$. Opsahl, Agneessens et al. (2010) proposes that the degree centrality is thus expressed as:

$$k_i = k_i \cdot \left(\frac{s_{ij}}{k_i}\right)^{\alpha} \tag{4.9}$$

Opsahl, Agneessens et al. (2010) also point out that the betweenness and closeness centralities are both addressed by defining the shortest path as the distance (where $distance = 1/weight$), whilst the Eigenvector centrality can largely be calculated in much the same way (although more robust algorithms that do not explicitly require the Perron-Frobenius theorem to be met are preferred – e.g. PageRank (Katz and Martin 1997, Page, Brin et al. 1999)).

Opsahl and Panzarasa (2009) address the clustering of weighted networks by considering the total weighted value of connections that form triangles divided by the total weight of the connections that form triplets as per equation (4.10).

$$C_{w_i} = \frac{value\ of\ triangles}{value\ of\ triplets} \tag{4.10}$$

Many of the other weightless measures are mathemtically still valid. However, the weighting changes the overall outcome significantly. In some cases, it improves certain approaches (e.g. partitioning is provided with greater discriminatory capabilities (Farine 2014), whilst in other cases, it provides less information (e.g. identifying tree structures in weighted networks is extremely difficult due to the changes in path lengths).

4.2. Social Network Analysis Review

Having defined a brief overview of networks, a lens through which SNA can be reviewed with regards to IDR has been developed. However, there are few SNA studies that have explicitly focused on IDR, and it is necessary to expand to closely related and analogous studies. The terms reviewed here are: “research”, “IDR”, “interdisciplinary research”, “knowledge management”, “cross-functional”, and “innovation” in conjunction with “networks” and “social network analysis”. These search terms were used in Google Scholar over all time periods to establish seminal papers and were then refined with more recent papers starting usually from 2013, but in certain cases where few relevant papers were found, 2004 papers onward were used. This creates a vast body of literature where certain concepts, phenomena, and methods occur multiple times.

SNA has existed for a long time, with (Moreno and Jennings 1934) pioneering the field. SNA has since been used to investigate a variety of different aspects of society. For instance, “The Small-World Problem” by (Milgram 1967), which concluded that people in the USA were separated by “six-degrees of separation”. Padgett and Ansell (1993) used networks to show that the Medici family were the most influential (central) family in Renaissance Florence through marriage, partnership, bank employment, trade, and real estate deals (Padgett and Ansell 1993).

With regards to IDR, five categories of SNA studies have been found and compiled under the following headings: “Knowledge creation cross-sectional studies”, “Research collaboration cross-sectional studies”, “Research collaboration structural studies”, “Research collaboration structural dynamics”, and “Knowledge creation structural studies and dynamics”.

4.2.1. Knowledge creation cross-sectional studies

The use of SNA to investigate organisational efficiency and innovative capability drew a considerable amount of research effort in the 1970s and 1980s (Steve Conway and Steward 2009). In this time, seminal works established many SNA concepts that are commonly used today.

4.2.1.1. Strength of weak ties

Granovetter (1973) is one of the most cited papers in the social sciences with over 46,000 citations (as per Google Scholar on 03.01.2018) and established the concept of ‘the strength of weak ties’ (Conway and Steward 2009). The work is based on the concept that strong ties are generated on “homophilization” of source-receiver knowledge, whilst weak ties are best characterised by heterophilous knowledge, and are therefore valuable (Rogers and Bhowmik 1970). It is important to note that Granovetter (1973) does not define how a strong or weak tie should be measured, although most works citing the study take it to mean that a strong tie is characterized by frequent interactions among members (Freeman, White et al. 1992). Granovetter (1973) suggests that effective communication leads to greater homophily in knowledge, and therefore heterogeneous ties provide better diffusion. Granovetter (1973) concluded that the greatest increase in path length is when weak-ties are cut, as these serve as bridges between different communities. However, no studies could be found where specific bridges are found, and most real networks seldom have outliers in betweenness scores (Barabási and Pósfai 2016).

Whilst it is a central concept in SNA, the results have been mixed. Weak ties and structural holes have benefitted the innovative capability (Perry-Smith 2006). Zhou, Shin et al. (2009) largely concurs, but suggests an inverted U-shape relationship. However, many papers find that strong ties are better at sharing knowledge, thereby reducing the cost of knowledge transactions between individuals (Kachra and White 2008, Phelps, Heidl et al. 2012). However, the observation that heterophilous knowledge is conducive to both spreading and creating knowledge has been corroborated by several different studies (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009, Backmann, Hoegl et al. 2015, Guan and Liu 2016).

4.2.1.2. Network centrality

Knowledge creation is a process driven by individuals interacting with knowledge artefacts in themselves, other people, and through other communication mediums (Nonaka, Byosiere et al. 1994, Phelps, Heidl et al. 2012). Combining knowledge from different fields in novel ways has been shown to be conducive to developing knowledge (Burt 2004, Nelson 2009). This has been the documented reasoning behind the success of many organisations such as IDEO (Hargadon and Sutton 1997). It is the ability to transfer, understand, and create valuable synergy from these different knowledge bases that determines the success of this process (Nahapiet and Ghoshal 1998). It is through the success of such mechanisms that investigating the effect of interpersonal relationships and the development of knowledge becomes vital (Singh and Fleming 2010). Within organisations direct ties communication has been identified as being vital to sharing more reliable and more complex information (Singh 2005). Knowledge networks have been studied in many different contexts: diffusion of knowledge (Bothner 2003, Nerkar and Paruchuri 2005), knowledge production (Jackson 2010), team knowledge exchange and creation capabilities (Reagans and McEvily 2003), and interorganizational strategic alliances to improve knowledge transfer and innovation (Lane and Lubatkin 1998, Schilling and Phelps 2007).

However, where many knowledge networks findings show that there is a positive relationship between innovative capability of an individual and the number of links that person has (Audia and Goncalo 2007), other studies claim that the number of ties is negatively correlated to the quality of their research (Bordons, Aparicio et al. 2015). McFadyen and Cannella (2004) investigated a network of biomedical research scientists and found that both the number of social relations and the strength of the interpersonal relations had diminishing returns on knowledge creation (measured as the sum of journal impact factors researchers submit to and normalised by the number of co-authors). This is supported where a very similar correlation was found and attributed who correlated the collaborations network to exploitation and exploration innovative capability (Guan and Liu 2016). Both studies have found an inverted U-shaped correlation.

McFadyen and Cannella (2005) investigated the impact of geographic proximity and department interdepartmental collaboration and concluded that geographic proximity is not as strong a predictor as department; they found that the further the collaboration moves away from their own department, the greater the knowledge produced (measured as the sum of journal impact factors researchers submit to and normalised by the number of co-authors).

Guimera, Uzzi et al. (2005) establishes that there is generally a statistically significant positive correlation with the impact factor and the probability of including existing members, and negative correlation with the impact factor and the probability of selecting past collaborators. This suggests

that introducing new knowledge into a network is conducive to knowledge creation and academic performance.

Some studies have used the centralities to classify nodes into organisational archetypes (e.g. by identifying stars, liaisons, gatekeepers) (Tichy, Tushman et al. 1979). Such an approach has been revisited more recently in Batallas and Yassine (2006) who propose brokerage indexes for the archetypes, which can be used to alter management strategies to suit the needs of the situation.

4.2.1.3. Network clustering

Other studies have investigated the local structure of the network, identifying the lack of local cluster (open triads) is associated with greater innovative capability, as it is associated with a structural hole indicating that diverse ideas flow to the person (Nerkar and Paruchuri 2005). Such approaches suggest that knowledge diversity is greater if their neighbours are not linked.

However, other studies have found that closed triads perform better with greater flow of idea between the individuals who can then collaborate, a concept close to IDR (Obstfeld 2015). McFadyen, Semadeni et al. (2009) offers a different perspective: researchers maintaining strong ties with researchers with a sparse ego-network provide the knowledge creation.

Guan and Liu investigate exploitative and exploratory innovation in both knowledge and collaboration networks, testing six hypotheses on each (Guan and Liu 2016). The main findings were that stronger integration across clusters provide greater capability for knowledge diffusion.

There have been conflicting findings of high local clustering with regards to tie strength. Pan and Saramäki (2012) found that high local clustering were associated with weak ties, whilst the opposite has also been found (Uddin, Hossain et al. 2013). Degree and betweenness centralities were found to increase the citation count, and were central to developing strong ties, and that these were found in nodes with high local clustering (Uddin, Hossain et al. 2013).

4.2.1.4. Other knowledge diffusion methods

Some studies have focused on the effect that the average path length has on knowledge creation, finding that shorter path lengths increase knowledge transference and innovation performance (Fleming, King et al. 2007).

Other approaches have not been specific to scientific collaborations but provide analogous findings. In the fields of rumour propagation, Banerjee, Chandrasekhar et al. (2014) defined the geographic strength as 1/distance to find geographic centrality measures, to be correlated to the diffusion of gossip. It was found that a new diffusion centrality matched real events well. In the field of social

capital, Ellison, Vitak et al. (2014) investigate the importance of maintaining relationships in order to develop social capital. The main finding of the paper was that merely being connected to other members in social networking sites did not produce social capital, but rather that the small efforts to maintain specific relationships did, making frequency weighting vital.

4.2.2. Research networks and citation indices

There is a growing body of SNA literature regarding how it is that networks affect output. For instance, it has been found that the amount of collaboration and co-authorship has been found to be rising overall with time (Luukkonen, Persson et al. 1992, Luukkonen, Tijssen et al. 1993, Kronegger, Ferligoj et al. 2011). Other methods have focused on identifying links between network position and academic output.

Collaboration networks have yielded many studies and have reported to providing many of the same benefits that IDR offers: improved access to expertise for complex problems (Katz and Martin 1997, Sonnenwald 2007, Hale 2012, Cimenler, Reeves et al. 2014), growth in academia due to cross-fertilization (Cummings and Kiesler 2005), increased human capital development (Bozeman and Corley 2004), and increased productivity (Lee and Bozeman 2005). Ye, Li et al. (2013) identified two types of researchers, unifying researchers who establish new connections, and researchers who strengthen their existing relationships.

Centralities have been postulated as corresponding to different social processes (Freeman 1979, Freeman 1980), which impact on these increasing levels of collaboration. Freeman (1978) postulated that degree centrality measured the extent of communication (Freeman 1978), whereas betweenness centrality represented the ability for an individual to control information propagation (Ye, Li et al. 2013). Ye et al. go on to show there was a correlation between the centralities (degree and betweenness) and academic productivity (Ye, Li et al. 2013).

Li, Liao et al. (2013) approach co-authorship networks using a social capital lens. They use Nahapiet and Ghoshal's (1998) approach to social capital in research, and establish three dimensions of social capital as being important: structural (e.g. centrality), relational (e.g. trust), and cognitive (e.g. knowledge) (Nahapiet and Ghoshal 1998, Li, Liao et al. 2013). Structurally, degree, closeness, and betweenness centralities correlate positively with higher academic outputs. They also find that collaborating with higher output academics correlates positively with academic outputs, and that scholars with diverse collaborators and longer tenure produce positive correlations. Furthermore, they establish that all these measures are positively correlated with one-another, suggesting that networks can be used to predict all three dimensions (Li, Liao et al. 2013).

To further explore the use of these centralities, several studies have investigated the influence of G- and H-indexes (for definitions see Chapter 3).

Abbasi, Chung et al. (2012) establishes that both degree centrality and betweenness centrality correlate positively to researchers' G-index. The paper also establishes that the larger a person's degree is compared to their neighbours, the more efficient they are deemed, which is positively correlated to the G-index. Bordons, Aparicio et al. (2015) also shows that there is a positive correlation between the degree centralities and the G-index, although this is less pronounced in the field of statistics. This alteration in the field of statistics, compared with their other fields of study, demonstrated that structure of statistics is sparser and more fragmented than the other two fields was hypothesised that this was due to the theoretical nature of statistics. This suggests that the correlation of degree and G-index is dependent on the density or fragmentation of the field's structure and lack of structural holes (i.e. dense networks) correlate negatively to performance (Abbasi, Chung et al. 2012). Loose field structure may also be indicative of low maturity of the field, for example within tourism and hospitality (Ye, Li et al. 2013).

Local clustering has been found to have a positive correlation with the G-index in Cimenler, Reeves et al. (2014), where they investigated the correlation of the network position and the G-index in the University of South Florida's College of Engineering (Cimenler, Reeves et al. 2014). They found that the number of collaborators, repeat collaborations, and redundancy (high local clustering) in connections to well-connected groups correlated positively with the G-index, whereas the eigenvector centrality correlated negatively. Conversely, Abbasi et al. found that higher local clustering is negatively correlated to the G-index (Abbasi, Chung et al. 2012). In an earlier paper by the same author (Abbasi, Altmann et al. 2011) found that the eigenvector centrality had a negative correlation with the G-index (Cimenler, Reeves et al. 2014). This disparity was reasoned to be due to Cimenler, Reeves et al. (2014) excluding students and external collaborators, whereas Abbasi, Altmann et al. (2011) find that well-performing professors have lower than expected eigenvector centralities due to them supervising many low eigenvector centrality students. The two papers are thus in agreement and provide a causal link as a control variable is not present in one of the groups.

With regards to high local clustering and the H-index, a negative correlation has been found (Hâncean and Perc 2016). Hâncean and Perc (2016) investigate the 'Matthew effect' (whereby known academics receive higher citations) in east European sociologists. They find that there is a positive correlation between the H-index and the mean H-index of neighbours. They also find that the betweenness centrality and the network size have a positive correlation with the H-index. They also agree with major findings that high local clustering has a negative correlation with the H-index.

However, it has been shown (Barrat, Barthelemy et al. 2004) that the average weight of ties is not dependent on the degrees of the pairs, indicating that additional importance is not necessarily placed on repeat collaborations with highly connected nodes, suggesting that the Matthew effect either does not apply, or does not apply to repeat collaborations. They found for the Scientific Collaboration Network that the exponent was close to 1, indicating that $s(k) = k\langle w_{ij}\rangle$, and there was no overall preference to collaborating with old acquaintances, suggesting a fluid scientific structure (Barrat, Barthelemy et al. 2004).

Subsequent studies on scientific network structures are sparser and focus more on correlating ego-centric measures to output as cross-functional studies. The development of the H-index sparked an interest in trying to establish an academic network centrality measure.

Zhao, Rousseau et al. (2011) propose a H-degree based off the academic H-index (measured as the largest number of links, *H*, with at least weight *H*). Yan, Zhai et al. (2013) extends on this work to develop a C-index that can be used as a centrality measure to the collaboration competence. The collaboration competence is defined as the H-index of the product of the edge strength and the neighbours' C-index, giving it an eigenvector centrality element. The algorithm is shown to be stable and to yield a scale-free distribution for scale-free networks (Yan, Zhai et al. 2013).

The H-index is related to the structure of a network and has been shown to be positively correlated with the degree, closeness, and betweenness centralities (De Stefano, Fuccella et al. 2013). The H-index also has a small positive correlation for greater external collaborations (De Stefano, Fuccella et al. 2013), which is a surprising result given that IDR suffers from a lack of appreciation (see Inhibitors to IDR and their associated costs). However, these findings were also based on the field of mathematics, which may have a differing dynamics as applied mathematics is interdisciplinary.

The density of networks has largely been suggested as providing a negative impact on academic performance due to the lack of structural holes (Sparrowe, Liden et al. 2001, Ortega 2014). However, other studies have shown that there is no statistically significant trend with degree, or betweenness centrality on the H-index (McCarty, Jawitz et al. 2013).

Subsequent decision-making information, through the prediction of highly cited articles can be achieved with machine learning (Sarigöl, Pfitzner et al. 2014). Using Machine Learning algorithms trained on networks centrality, they were able to predict with high accuracy whether an article would be highly cited. This was done on the basis that time-evolving collaboration networks and citation numbers provide an indication to a trend, showing that longitudinal networks, or network evolution plays a central role. Furthermore, they conclude that no single network centrality provides strong predictions, and a combination of centralities are needed. Specifically, degree, betweenness, and eigenvector centralities play an important role (Sarigöl, Pfitzner et al. 2014).

However, it is important to realise there are limitations to creating co-authorship networks. For instance, it has been reported that collaborations do not always result in co-authorship as authors may choose to publish in their own fields (Katz and Martin 1997, Cimenler, Reeves et al. 2014).

This provides a clear overview of the main body of work that is analogous to IDR.

4.2.3. Research network structures

This section reviews papers that review the overall state of the network structure and discusses its effects with regards to IDR.

Newman (2001) was an early pioneer utilising digital databases to construct networks, resulting in the first large scale SNA. Whilst the author outlined a difficulty in such data with multiple cases of a single author being listed under different naming formats, which may have an impact, they show that the number of papers per author follows a power-law (or truncated power-law in some cases), and that the number of authors per paper also follows a power-law (albeit with much higher exponents). With this, 80-90% of the network nodes are filled with a 'giant cluster' or 'giant component' and that certain fields have different local clustering values than others, indicating that these have different collaboration cultures or social dynamics. It should be born in mind that due to differing structures found throughout various fields, that IDR occurs more, for instance, between certain fields than in others (e.g. biological sciences was very interdisciplinary, whereas linguistics, letters, and arts was very interdisciplinary, but only with humanities) (Mena-Chalco, Digiampietri et al. 2014).

These collaboration dynamics are not unique to certain fields only but are also demonstrable among countries and regions.

SNA have also been used to map and outline collaboration within hospitality (Nunkoo, Gursoy et al. 2013) and tourism (Benckendorff and Zehrer 2013), where most of the research is centred in the developed world, but that this research bleeds into and benefits economies who rely on tourism to greater extent (Nunkoo, Gursoy et al. 2013). There is, however, some concern about fields drowning out the expertise in developing economies (Nunkoo, Gursoy et al. 2013).

Munoz, Queupil et al. (2016) found overall density of the network with Chile and Latin American is low and is regionally centred. Sparse and fragmented collaboration cause low levels of information propagation, and consequently, low scientific output. The creation of multiple paradigms may provide an initial barrier to collaboration between the fragmented groups but unifying the multiple groups may result in a burst of academic creativity (Ho, Nguyen et al. 2017).

The use of known effective collaboration structures (e.g. bridges, clusters, structural holes) should be used to improve scientific collaboration, which in turn can advance national capabilities through international collaboration (e.g. further collaborations with US institutions) (Mena-Chalco, Digiampietri et al. 2014). SNA can be used to identify how it is that research is occurring in the field of e-governments with respect to the triple helix mode. This provides a simple and effective way of measuring the research and interests in given fields (Khan and Park 2013). These methods can then be used to determine a country's overall R&D capability (Guan, Zuo et al. 2016) and as such, it is possible to use SNA to provide an overall indicated for R&D efficiency and performance.

The influence on epistemic direction on both a national and international level is very much determined by a clique of social scientists (Benckendorff and Zehrer 2013). For instance, (Linden, Barbosa et al. 2017) discuss a sentiment of unfairness in Brazilian academia with regards to publishing criteria. The authors suggest that this is not the case, but rather that the sentiment exists because there are epistemic differences in Latin American academia and publishing centres elsewhere. This is supported by network findings that show that the academic output is highly correlated with the position of the programs in the international setting, suggesting that greater international collaboration leads towards global paradigms (Linden, Barbosa et al. 2017).

Identifying overall structures have helped develop methods to identify thematic clusters within networks and provides additional information for policy-makers (Wu and Duan 2015, Mehmood, Choi et al. 2016).

This provides insight into how it is that overall network structures have been used to gain a deeper understanding of various systems.

4.2.4. Network dynamics simulations

The topology and its various statistical and analytical measures provide a great deal of knowledge regarding a system. However, the analyses provide mostly what could be considered macroscopic data. The microscopic abilities of individual Nodes are of great interest in SNA. Previous authors have been able to model both cascading of properties within networks (Carreras, Lynch et al. 2002, Motter and Lai 2002, Leskovec, Singh et al. 2006, Leskovec, McGlohon et al. 2007, Buldyrev, Parshani et al. 2010). Such a model could represent information flow. There have been many different attempts at modelling microscopic interactions, with varying degrees of success. The first thing that must be acknowledged is that in network dynamics there is a propagation of some property, which may or may not influence the topology of the system itself.

The most common approach to propagating a property is the Susceptible-Infected-Recovered (SIR) or Susceptible-Infected-Susceptible (SIS) from epidemiology, with varying degrees of success (May and Lloyd 2001, Barthélemy, Barrat et al. 2004, Volz 2008).

Pastor-Satorras and Vespignani (2001) focuses upon the SIS epidemic spreading in internet systems that can be carried through a variety of mediums (emails, FTP, etc…). These networks have been identified in previous works to be Scale-Free in nature (Barabási and Albert 1999).

Previous works on random graphs have found that there is an epidemic rate critical threshold, λ_c. Above this value the infection is persistent and below it the epidemic dies out (Kephart 1994, Marro and Dickman 2005).

Based on empirical data however, it has been found that the P(t) follows a power-law distribution. Furthermore, whilst most viruses die out within the first two months, many viruses survive for much longer periods. Seeing as anti-viruses respond within days or weeks of the first reported incident, this is indeed a significantly long survival period.

In simulating this, the Barabási-Albert algorithm is implemented for N=10^3 to $8.5\text{x}10^5$ Nodes and half the Nodes are initially infected. What is most interesting about the results is that in Scale-Free networks, there is an absence of the critical rate threshold (viruses are prevalent), whereas in bounded networks they have a clear threshold. In fact, it was found that the steady-state density is independent of the network size. Conversely, the time to reach said steady-state is not.

It is assumed when the exponent of scale-free graphs reaches 4 (exponential tails), then the re-emergence of a threshold will exist. This seems to be confirmed in another study by the same authors.

It is an effective model that is significantly affected by the topology of the network, though it relies on states as opposed to propagation and development of a property. **A**nalysis of the effect of perturbations by analysing the spectral density (De Aguiar and Bar-Yam 2005) demonstrate that not only does the density of states contain information regarding the topology, but also of the dynamics to external perturbations.

Other approaches vary significantly, and researchers have stated the inability to find commonality between the various papers (Barzel and Barabasi 2013). However, by adopting a framework that mimics network dynamics, a universality in network dynamics can be found. Barzel and Barabasi (2013) develop such a model based on two terms, a term that develops the property on the basis of the property itself. The second term is how the property in one Node is affected by the properties in the neighbouring Nodes.

$$\frac{dx_i}{dt} = W\left(x_i(t)\right) + \sum_{j=1}^{N-1} A_{ij} Q(x_i(t), x_j(t)) \quad (4.11)$$

Where *W* is a function concerned with altering itself, *Q* is a function that concerns the effect of neighbouring Nodes, and A is the adjacency matrix. Using this form, four separate studies using different dynamical methods are adapted to this form.

The models that are investigated are a Biomechanical model - B (Mass-Action Kinetics model) (Voit and Radivoyevitch 2000), a Birth-Death model - BD (Population Dynamics model) (Hayes and Babu 2004), a Regulatory dynamics of gene regulation – R (Michaelis-Menten model) (Karlebach and Shamir 2008), and an epidemiology model - E (SIS) (Pastor-Satorras and Vespignani 2001, Hufnagel, Brockmann et al. 2004, Dodds and Watts 2005).

Burstiness is a concept in Network Dynamics that the property propagation does not occur with a smooth distribution, but rather occurs in sporadic and intense bursts. This has been studied by Barabasi (2005) who analyses the emergence of heavy-tails in human dynamics. The system-wide dynamics of many systems (e.g. social, technological and economic) are driven by human dynamics. This puts understanding human behaviour at the centre of many real-world challenges. Several models for human behaviour has already been predicted by Poisson processes (Haight 1967). However, there is increasing evidence that in the context of work patterns, communication and entertainment that this is not the case. In fact, it seems to be characterised by burst of activity separated with long periods of inactivity (Barabasi 2005, Karsai, Kaski et al. 2012).

Barabasi (2005) attributes this burst activity to be a consequence of decision-based queuing.

This provides a strong insight into how it is that SNA has been used to study human processes.

4.3. Chapter summary

Having reviewed the SNA literature, no studies could be found that provided direct models to identifying individuals who enable and sustain IDR. However, many models could be found that are analogous, or with relatively few tweaks could be made to identify such individuals.

Five main models were identified in the literature:

1. Degree centrality
2. Betweenness centrality
3. Eigenvector centrality
4. Structural holes (clustering)

5. The strength of weak ties

Furthermore, many papers focused on citation-based output metrics as a way of determining whether research was suitable.

These provide the foundations of the next steps of the research. According to the adopted research methodology, the next steps must construct an instrument for data collection, select a sample, write a research proposal, and collect the data.

Chapter 5: The University of Bath Co-Authorship Networks

This chapter establishes the instrument for data collection, the selection of a sample, and the actual collection of the dataset as per the adopted research methodology (see Chapter 2).

Having established that collaboration networks analysed under the paradigm of SNA have been used as a robust approach to investigate research organisations, team dynamics, and knowledge dissemination, it is necessary to define a dataset that can adequately perform similar analyses. To determine what constitutes a suitable dataset, the following questions were answered.

1. What data does a collaboration network need?
2. What are the research boundaries?
3. What are the requirements of the data to make it fit-for-purpose?
4. What is a suitable network data source?
5. What success metrics are suitable?

Having established these, a suitable sample is proposed, a data collection instrument is constructed, and the data is collected.

5.1. Collaboration network

Collaboration networks are the product of constructing a network based on all the collaborations occurring between a set of individuals.

At a fundamental level, the network needs to be able to construct all the basic measures that have been established in literature (see Chapter 4). This requires a list of nodes and a list of links, which represents researchers and collaborations respectively. A collaboration between two researchers can be represented by a binary representation, 0 or 1 representing no existing and existing collaboration respectively. This could be used to build an adjacency matrix as outlined in Chapter 4. Further information can be collected to identify weight of collaborations (e.g. frequency), or the time of collaborations.

Additionally, in order to capture IDR collaborations, it is necessary to identify disciplines within the network. This is established in the next chapter (Chapter 6).

5.2. Organisational boundaries

Having defined the elements needed to construct a collaboration network, it is necessary to identify the boundaries of the network so that reflective data can be collected. As the purpose is to identify individuals, it must be from a set of individuals. The most natural sets are made from organisations

(e.g. teams, departments, universities). Therefore, the research boundaries are set within research organisations, and the network should reflect this. As the research sought to benefit decision and policy makers – two major stakeholders were identified in this research: University senior management and research councils. Universities as the research organisation has a good overlap for both and therefore chosen as the research boundaries.

This is not to say that collaborations between organisations are not important, they are without a doubt, but a practical boundary is needed. However, given the resources, this research should expand the boundaries to include interorganisational collaborations, which will likely be vital. It is therefore important that the research remains extensible to other (similar) datasets.

The University of Bath is an exemplar research organisation. It was chosen as due to it being embedded within the University and interacting with other individuals embedded in it provides a face validity.

It is important to discuss several aspects of choosing this as the boundaries of the research.

First, it is important to note that this is not strictly speaking convenience sampling. Convenience sampling does not create a random sample of a population, thereby introducing a bias. By sampling the entire University of Bath, there is indeed a bias, but not one generated by the researcher or by the data collection method. Instead, the bias is a result of the boundaries chosen, and the bias would exist for any organisation. That is to say, if a different organisation would have been chosen, the same bias would exist. This implies two things: that the University is an exemplar organisation no different than other universities, and within the University, the entire population is sampled, and is therefore a good representation of the various individuals. Furthermore, with over 2000 published researchers, the population size is large enough to establish statistical significance over many regressions. Ultimately, this implies that analyses are only corroborated to the specific University, and further validation will be necessary to make the work truly extensible. This represents vital further work.

Second, in UK universities, REF dissuades researchers from collaborating with members from the same department as only one researcher can claim credit for a paper. There is therefore the question of how this affects the collaboration within University. Does this reduce the amount of disciplinary collaboration, or does it only reduce the amount of disciplinary co-authorship? These are research questions that require sociological approaches to answer and are outside the scope of the research, but will nevertheless affect the analysis. It is therefore important to discuss the various possibilities and how it might affect the research. There are four possible cases:

- REF significantly dissuades researchers from collaborating with disciplinary colleagues and significantly less collaboration occurs.

- REF significantly dissuades researchers from collaborating with disciplinary colleagues, but some collaboration occurs regardless.
- REF significantly dissuades researchers from collaborating with disciplinary colleagues, but collaborations occurs regardless, albeit with fewer published papers.
- REF dissuades researchers from collaborating with disciplinary colleagues, but it does not affect collaborative patterns.

In the first case, if less collaboration occurs with disciplinary colleagues, then it stands to reason that co-authorship will capture this.

In the second case, if collaboration occurs regardless, then co-authorship will capture this.

In the third case, co-authorship will not accurately capture disciplinary research.

In the fourth case, if REF has no effect on collaborative patterns, then co-authorship networks has no additional patterns to capture.

Therefore, only the third case will affect co-authorship networks. However, given that the vast majority of publications are disciplinary (see Chapters 7 and 9) from 2014 (when REF was implemented) onwards, either the barriers to conducting IDR are even greater or the disincentive to publish together is less than might be thought (perhaps all disciplinary authors collaborate, write separate papers, and name other authors as non-primary authors). Exploring the data, no significant changes can be seen from 2014 onwards in comparison to the periods before it, suggesting that this is not a significant issue.

A third possibility is that REF dissuades researchers from collaborating with colleagues within the same University, and thereby encourages inter-organisational disciplinary collaboration to occur. However, this case is outside the scope of the research, but remains a very important area for further research.

5.3. Data requirements

It was vital to collect data from a robust and as unbiased a source as can be. Data fidelity is arguably the most vital component of any study. The well-known phrase “Garbage in, garbage out” applies to any input-output system. Therefore, if any analysis is to be trusted, or any model is to be created, it is vital to ensure that these are based on high fidelity data, or that the limitations of the data are understood (provided it is still suitable for its purposes).

To create such a dataset, the data requirements in this research were determined by four different aspects.

1. **The data boundaries need to match the research boundaries.**

The chosen boundaries in this research are organisational and for the dataset to be pertinent, it must match these boundaries.

2. **The data source needs to represent the research focus as accurately and objectively as possible.**

The focus in this research is IDR. Therefore, the data needs to establish the disciplines of individuals, and whether they have collaborated in research.

The most objective and automatable data source is official collaboration documentation. This circumvents most issues regarding human biases in collecting the data in the first place, such as designing and answering surveys. Based on the network requirements, the data needs to be able to establish links between known persons, the weight of these links, and what discipline people work in. One of academia's main outputs is in publishing scientific papers, which includes the authors. This provides a collaborative link between researchers that can be used to construct a network. Furthermore, the number of papers authors have co-authored can provide the weight of such links.

If collaborations were restricted to co-authorship in journal publications, it would be an indication that a significant collaboration has occurred.

3. **The data must be quantitative.**

This stems from two different factors: the deductive approach chosen for this research and the gap quantitative studies on IDR. SNA is inherently a quantitative method.

4. **The data source should be openly available and procured from an objective and unbiased source.**

One further criterion is proposed for repeatability, expansion, and maintenance. This is that the data should be easily available. Taking advantage of big data availability allows elimination of human sources of bias in quantitative methods such as observation or surveys, which have been shown to cause several problems (Hamill 2006). For instance, samples from a population has been shown to not produce the same scale-free distribution the full population does (Hamill 2006, Newman 2010). This means that the Central Limit Theorem used in cross-sectional population studies cannot be applied to determine the overall structure (Mendenhall and Sincich 2016). Other methods, such as snowball sampling, have boundary issues (Newman 2010). Furthermore, even if these fundamental issues were overcome, surveys suffer from response percentages (which are particularly troublesome for networks as these tend to be the peripheral nodes), and shallow depth answers (e.g. a highly connected node is unlikely to list more than eight collaborators, are prone to forgetting details, and it is difficult to codify written responses) (Wasserman and Faust 1994). This is in

addition to biases in responses (e.g. the type of people who give lengthy responses might skew the data).

Based on these requirements, a journal-based co-authorship network would provide a suitable proxy for collaboration networks. Journals are preferred over other collaboration publications as it is assumed that a significant amount of work has gone into producing journal articles, and it is best not to include other works which may been the result of less work (e.g. conference papers).

5.4. Data source

Based on the requirements outlined above, a journal co-authorship network in the University of Bath from 2000-2017 served the purposes of this research. This was chosen specifically as this data was readily available and can easily be reproduced in different universities using the same method.

Co-authorship networks centred at the University of Bath meet all the criteria required, as outlined in Table 5.1. The official collaboration documentation can be found on http://www.opus.bath.ac.uk (accessed 11/08/2017 at 17:37) This data source is determined to be suitable as several different universities maintain similar databases (e.g. University of Cranfield, University of Bristol), albeit with slightly different structures.

Other sources are possible such as Open Researcher and Contributor ID (ORCiD) or Gateway to Research (GtR) are possible. However, these both suffer from major issues. ORCiD is maintained by individuals and their data is not publicly available and therefore cannot be scraped. It also suffers from the fact that it must be updated by users and will therefore be likely to be a very incomplete set. Furthermore, it does not allow individuals to be searched by institutions, thereby requiring persons of interest to be known before the data can be collected.

GtR will only contain publication data that has been officially affiliated with a grant. Therefore, any research which is not directly related to a grant is not included.

Both of these do not provide an easy way to identify Bath authors, thereby making it difficult to establish who should and should not be included in the data given the research boundaries.

Therefore, the data source is deemed to be best publicly available dataset, and provided that it is accurately maintained by the University, will likely to be highly accurate without issues of mistaken identities. However, with access to data from organisations such as Researchfish that collect more comprehensive details about research in the UK, this research could easily be extensible to multiple universities and include interorganisational collaborations.

Table 5.1 – Co-authorship networks eligibility.

Requirements	Fulfilment
Able to build a network capable of creating the outlined measures.	Co-authorship networks are able to associate links and number of collaborations between two individuals. Furthermore, such data can be associated with a discipline.
1. The data boundaries need to match the research boundaries.	Coauthorship networks would match these provided that a complete set of papers can be found for the desired boundaries. In the case of this research, around research organisations. The University of Bath provides a list of publications and their details online: http://www.opus.bath.ac.uk .
2. The data source needs to accurately represent the research focus.	The co-authorship network would in most cases show a direct collaboration between two individuals. No data is perfect however, and it must be taken into consideration that a few false positives will exist.
3. The data must be quantitative.	The use of the networks paradigms ensures a quantitative approach.
4. The data source should be openly available and procured from an objective and unbiased source.	The data is openly available in many cases (e.g. http://www.opus.bath.ac.uk). The use of co-authorship networks is largely unbiased. Only two sources of human uncertainty could be included at any stage: the inclusion of authors on a paper, and any omission of published works from the data source.

A few errors and biases exist in this dataset. For instance, if any publications are overlooked, this could potentially skew the findings. Furthermore, co-authorship credits are not always the results

of direct collaborations, with many people included as they share a grant for instance. In other cases, people who have collaborated, are not included.

As journal papers are the result of significant work, it is assumed in this research that the individuals collaborated. Given the quantitative nature of this work, and that networks provide cross-sectional results, it is likely that any issues will be dampened by the statistical nature of the data.

5.5. Instrument of collection

Two avenues to collect data from the University of Bath opus (http://www.opus.bath.ac.uk) were possible. The first was to request the data from the University itself. The second was to collect the data from the website as it was open source.

As there is a desire to keep the data collection method repeatable and cheap, crawling the website to collect the data is preferable. The website's robots.txt (a document every website has that outline the terms of crawling a website) allows for crawling without restriction (as of the time of writing).

A scraper is a type of web-crawler that can follow hyperlinks and automatically extract desired data. Unlike a crawler, it is highly targeted and is not used to explore the world-wide web, but to pull out all pre-defined data structures.

A Python 3 implementation of a scraper, 'Scrapy', provided a powerful framework to perform this scraping whilst being mindful of the server traffic. It allows data structure to be extracted using extended CSS selectors.

The scraper crawls every person's site, finds that person's publications, and from the publication site, the following information is stored in a data structure.

1. The authors
2. The bath authors
3. The bath authors unique ID
4. Date of publication
5. Journal published in
6. Department published in
7. Centre published in
8. Publication type

Names are given a format of < "last name", "first name" > and titles are removed from the name.

The time boundaries were set to include all significant digitization of records. As such, all papers from 2000 to 2017 were recorded. This included 38,173 publications. 23,468 of these were journal

publications. 16,292 journal publications had an abstract associated with them. 2,775 unique University of Bath authors and 39,975 authors total were found.

This already shows that the vast majority of collaborations are outside of the boundaries set and is a severe limitation. It is, however, impossible to say how many of the external authors are caused by different spellings of the name.

As this data is grouped by publication, any author who appears in this publication will be a co-author. This establishes links to create an adjacency matrix, where every instance two authors appear on the same publication, a 1 is recorded, and otherwise remains 0. The weight of the links is found by the number of times they appear together on publications.

Each publication is time stamped, so a time frame can be chosen to include or exclude a publication.

5.6. Data validation

Having used the instrument of data collection to create the dataset, it is necessary to determine the validity of the dataset. However, for broad longitudinal data, this is difficult to do in a bespoke way. As such, the data are validated by comparison to other similar datasets. There are two metrics of validation which are addressed by the following hypotheses.

Hypothesis 5.1 – The University of Bath co-authorship network exhibits a scale-free degree distribution with an exponent between -2 and -3 (Newman 2001, Albert and Barabási 2002, Barrat, Barthelemy et al. 2004).

Hypothesis 5.2 – The University of Bath co-authorship network shows that the average number of collaborators is increasing faster than the number of papers (Luukkonen, Persson et al. 1992, Luukkonen, Tijssen et al. 1993).

These are both corroborated as can be seen in Figure 5.1 and Figure 5.2. As this data corroborates findings from other studies (see references associated with Hypotheses 5.1 and 5.2 respectively), it is considered a validated dataset.

It is worth noting that validation by comparison is entirely dependent on the original study being correct. Strictly speaking, the dataset is not validated, but simply corroborates the findings from the original paper. However, as there is no way to falsify the dataset, this is the best possible ‘validation’ that can be achieved.

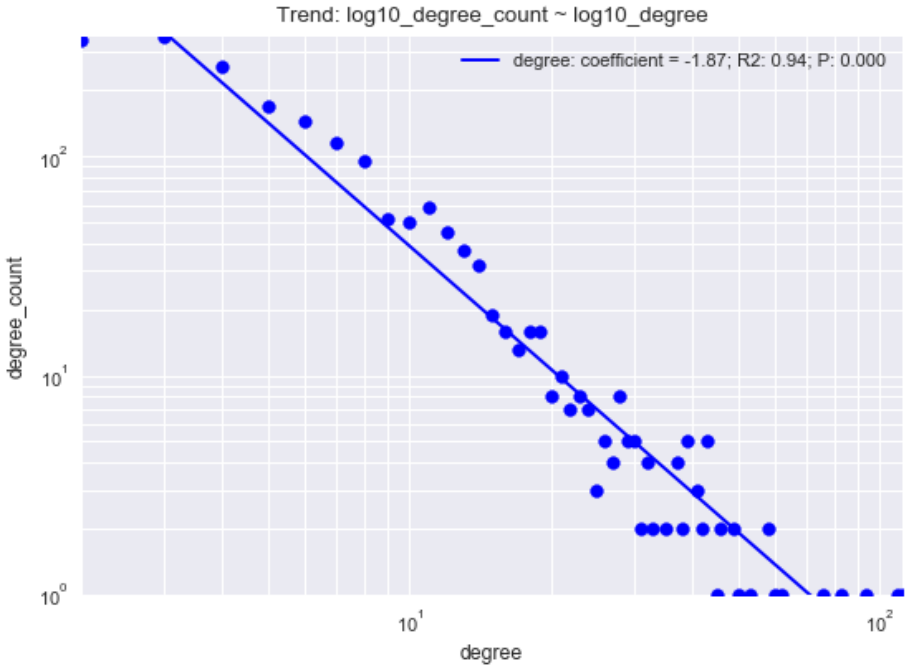

Figure 5.1. Degree distribution exhibiting typical scale-free behaviour.

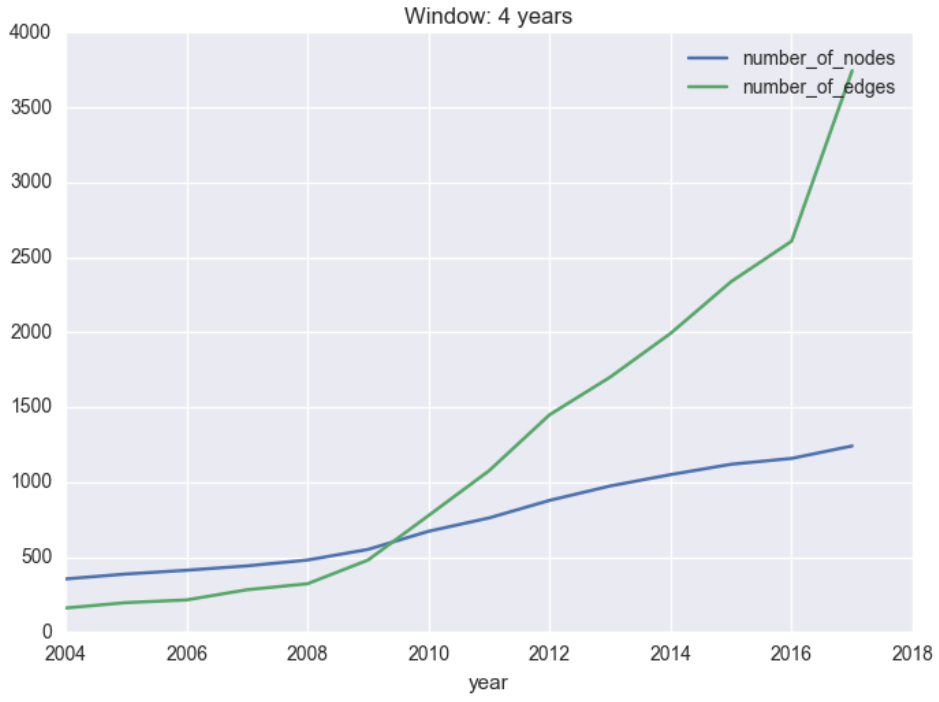

Figure 5.2. A figure showing the number of authors and the number of links occurring in a 4-year period (i.e. 2000-2004 to 2014-2018). The number of links is increasing exponentially faster than the number of authors.

5.7. Metrics of success

Having created a network, it was possible to create statistical measures to see what an author's network position was. However, this gives no indication as to whether an author will be successful at enabling or sustaining IDR.

This section proposes three different operational definitions metrics of success: bibliographic measures, funding, and future connectivity.

5.7.1. Bibliographic measures

Bibliographic measures are defined here as a measure of the quality and quantity of research outputs in the form of peer-reviewed publications. This is the main metric used in literature (see Chapter 4). It is reasoned that capable individuals who create high quality work would attract and enable future collaborations in intra and interdisciplinary research. This provides a robust measure of success as peer-reviewed publications are the main medium for sharing the academics' work. However, the issue with bibliographic measures is that most are based on the number of citations, which is entirely dependent on the popularity of the subject matter and whether a particular topic is á-la-mode. This skews certain topics very significantly and makes it very difficult to compare researchers between and even within disciplines.

Commonly used measures in academia include the H-index. However, such data is not easily available, and can be quite expensive to find historic values, and in bulk. It is easier to find data on specific articles that would allow these values to be calculated in bespoke timeframes. However, even the number of citations proved difficult to find, and had a great cost associated with them.

The easiest bibliographic measure obtainable is simply the number of papers produced, given in equation 5.1.

$$Y_{papers_i} = K \tag{5.1}$$

Where the bibliographic measure based on the number of papers, Y_{papers_i}, is simply the number of papers, K, author i appears in.

Another relatively easy measure to find is the impact factor of journals. However, many different databases claim variations of impact factors for the same journal. Based on reputation, this research uses the Thomson-Reuters impact factor values for 2015 outlining the impact factors of 12,870 journals. It is important to note that the impact factors are static, and unless a paper was published in 2015, the impact factor is wrong. However, the impact factor values do not change much over

time, and these values were deemed fit-for-purpose. Future work to ensure that these values do not change significantly is necessary.

Having identified the impact factor of a paper by identifying its journal, it is important to understand that the impact factors need to be associated with both a link and a node. This can be done by simply summing impact factors across given co-authorships and authors respectively. These are given below.

$$Y_{IF\,ij} = \sum_{k=1}^{K} IF_{ij}{}_{k} \tag{5.2}$$

$$Y_{IF\,i} = \sum_{j=1}^{N} B_{ij} \tag{5.3}$$

Where, $IF_{ij}{}_{k}$ is the impact factor of publication k out of K total between co-authors i and j. This gives us the bibliographic measure based on the impact factor for a tie, B_{ij}, and for a node, B_i.

It is then easy to distinguish between intra and interdisciplinary bibliographic measures by imposing a distinction.

$$Y_{IF\,intra\,ij} = \sum_{k=1,\ \boldsymbol{discipline(i)=discipline(j)}}^{K} IF_{ij}{}_{k} \tag{5.4}$$

$$Y_{IF\,inter\,ij} = \sum_{k=1,\ \boldsymbol{discipline(i)\neq discipline(j)}}^{K} IF_{ij}{}_{k} \tag{5.5}$$

This means that the various network measures can be tested against the bibliographic measures. By identifying network measures that are specific to IDR, it is possible to find which IDR measures were correlated with bibliographic output.

Furthermore, as the bibliographic measures are separated into B_{intra} and B_{inter}, it is possible to see how these measures affect bibliographic measures specific to intra or interdisciplinary research.

However, despite the simplicity of this method a significant challenge was found in the implementation. Namely, the journal names collected in the network data were not consistent. That is to say that capitalisation, spelling, spelling errors, and data structures errors meant that the initial

run had many non-matches. The journal names often included year of publication, had additional punctuation and information.

Therefore, it was necessary to clean up the data by:

- Splitting the journal name strings on punctuation, and keeping only the longest string
- Removing numbers
- Replacing '&' with 'and' in both the journal names and the Thomson-Reuters data.
- Remove all brackets from both
- All letters were made lower-case

The journal name can then be compared to the Thomson-Reuters data. The closest match from the Thomson-Reuters data was chosen based on a string comparison (greatest number of character matches). Two additional conditions were also implemented: if the string comparison was greater than 89% matches (i.e. if 89% or more of the characters matched), and if a non-empty journal name was found inside the closest string comparison match in the Thomson-Reuters data. If either of these conditions passed, a match was found. These criteria resulted in the following:

- 5,584 different spellings were found in the data.
- No false positives within 400, manually checked, matched journals. This method was therefore deemed to be suitable.

5.7.2. Funding

Funding is a complex measure to be used as a representation of quality or ability. In an ideal world, funding would go to the individuals who are most capable of answering the research questions the funding intends to answer. This is obviously not a linear process as research often finds unexpected results, particularly in real-world problems. By assuming that most of the funding has been allocated effectively, it is possible to use funding as a proxy of success. This provides a robust measure of success as the process of receiving a grant is based on many factors such as connectivity to relevant actors (within and outside the research organisation), previous performance, and proposal. It is arguably the most complete metric of success. However, there is also room for error, where funding for a particular field is subject to competition from different research organisations, and therefore suitable candidates for funding may not receive any. This may make it difficult for funding to provide continuous correlations. Funding may also favour the Matthew effect (prominent academics being more desirable) (Hâncean and Perc 2016).

RCUK funding is publicly available from the 'Gateway to Research' website (http://www.gtr.ac.uk). The use of their Application Programming Interface (API) makes it trivial to retrieve large amounts of data provided that the grant identification code is known.

These identification codes were downloaded into 'Comma Separated Values' (CSV) based on search queries. The entirety of their database was included if the 'University of Bath' was listed as either a research organisation, or if one of the investigators or listed staff was affiliated with the University of Bath. This enabled the data from every grant matching the criteria to be stored in a data structure (saved a JavaScript Object Notation, JSON). 782 grants were identified as being associated with the University of Bath from 01/01/2000-01/08/2017.

A wide variety of data was made available through the API. However, despite the huge potential in each of these fields to provide extremely powerful data to investigate many aspects of research, very few of these were consistently filled out. For instance, whilst certain publications were included in many grants, delving further into the grant shows that these were not the only publications from these same authors on the same dataset in the same time. This goes to show that there is data omitted probably for several reasons. However, the lack of consistency makes it very difficult to manipulate. Another example is the subject field that the grant touches on. These are filled but are given in various levels of aggregation. Some fill out generic disciplines such as 'Aerospace Engineering' whereas others fill in highly specific terms such as 'Laser Doppler Velocimetry'. Whilst these can be used to provide an idea what the grant was about, it makes it very difficult to compare.

For this reason, only fields that are uniformly filled out are considered. These are:

- The investigators' names
- The funding amounts
- Funding start date and end date

Even the investigators' names cause issues as they rarely match to other data and can differ from grant to grant. This requires a significant amount of work and oversight to ensure that investigators are properly matched.

This research draws a distinction between the total funding amount and funding per annum, with funding per annum being a better representation. Furthermore, the funding is split between the individuals on the grant. It might be possible to propose a split in funding that provides greater funding to the principal investigator, but without justification for a specific numeric split, the funding is simply allocated equally to all named investigators. Therefore, the funding is proposed to be calculated in (5.6).

$$F_i = \sum_{k=1}^{K} \frac{f_k}{N_k \cdot T_k} \tag{5.6}$$

Where the funding associated with an individual, F_i, is dependent on the funding (f), number of people (N), and the length of the grant (T) associated with grant, k, out of all grants, K.

This means that the various network measures can be tested against funding. By identifying network measures that are specific to IDR, it is possible to find which IDR measures is correlated with funding.

It is worth noting that the number of grant receivers is much smaller than the network. Therefore, for the regression, the people with no funding were not included as it severely skews the data.

Unlike the impact factor, this measure is not applied to every publication, but rather to every individual. Therefore, it was necessary to identify every University of Bath author and associate them with a name in the funding data. The names from the funding data were organised into < "last name", "first name" > format. However, there was no consistency in name format, meaning that often middle names were included, or only the first name was abbreviated, or both the first name and middle names were abbreviated, these were sometime separated by punctuation, and other times by space, and other times not separated at all. Given that some names were only two letters long, this really compounded the problem.

The similar names were provisionally matched based on a string comparison. Five different conditions were proposed, and if any one of them passed, the match was accepted:

1. If the matches were extremely similar name (i.e. similarity > 0.96).
2. If the names were very similar names (i.e. similarity > 0.9) AND the last names match exactly AND the initials match.
3. If the names were similar (i.e. similarity > 0.77) AND the last names match exactly AND the funding first name is an initial.
4. If the names were similar (i.e. similarity > 0.7) AND the last names match exactly AND the funding first name exists in the network first name or vice versa.
5. Or if a middle name is included, if any of the initials match

This provides very few false positives, which could be manually remedied by including exceptions to the rule. This means that every grant is now associated with a person from the authors list.

However, any statistical analysis correlating network measures to funding, even including only funded nodes shows statistically insignificant results. Table 5.2 shows the statistical results of a multivariable analysis (it is important to note that the variables are not uncorrelated, but have all

been included as this provides a non-linear model that should be able to approximate some correlation if some existed better than any one variable). The p-value is greater than the 0.05 threshold making any one model a worse fit than statistically insignificant.

Table 5.2. The statistical results of the fixed effects panel data analysis of the various structural measures that were identified as being important vs yearly funding from 2000-2010 to 2000-2017. The analysis shows that the results are statistically insignificant, and that no trend can be found between or within.

PanelOLS Estimation Summary

Dep. Variable:	delta_funding	R-squared:	0.0017
Estimator:	PanelOLS	R-squared (Between):	0.0301
No. Observations:	7384	R-squared (Within):	0.0017
Date:	Fri, Apr 27 2018	R-squared (Overall):	0.0238
Time:	19:33:13	Log-likelihood	-8.62e+04
Cov. Estimator:	Clustered		
		F-statistic:	2.1817
Entities:	923	P-value	0.0534
Avg Obs:	8.0000	Distribution:	F(5,6449)
Min Obs:	8.0000		
Max Obs:	8.0000	F-statistic (robust):	0.8495
		P-value	0.5144
Time periods:	8	Distribution:	F(5,6449)
Avg Obs:	923.00		
Min Obs:	923.00		
Max Obs:	923.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	1.052e+04	3264.6	3.2212	0.0013	4116.3	1.692e+04
degree	1595.7	1100.9	1.4494	0.1473	-562.47	3753.8
degree-squared	-1.3256	4.4050	-0.3009	0.7635	-9.9608	7.3097
betweenness	-0.0249	0.1680	-0.1484	0.8820	-0.3542	0.3043
pagerank	3.413e+06	5.928e+06	0.5758	0.5648	-8.208e+06	1.503e+07
structural_holes	-1174.1	686.82	-1.7094	0.0874	-2520.5	172.32

F-test for Poolability: 22.243
P-value: 0.0000
Distribution: F(929,6449)

Any hypothesis that uses funding as the dependent variable was rejected and is therefore unsuitable to be used a metrics of success. However, this in itself is valuable information and provides some indication that there is a disconnect between network structure and bibliographic measures to funding.

5.7.3. Future connectivity

The final measure of success is based on the wording of the research aim. A person who undertakes and sustains IDR is an individual who fosters additional IDR. Therefore, increase in degree from one period to another can be thought of as a measure of success.

$$Y_{k_i} = k_{future_i} - k_{present_i} = \sum_{j=1}^{N} \boldsymbol{A}_{future_{ij}} - \sum_{j=1}^{N} \boldsymbol{A}_{present_{ij}} \tag{5.7}$$

This means that the various network measures were testable against future co-authorships. It was therefore possible to compare IDR measures performed in comparison to the overall population, both in terms of overall performance and in future IDR. However, it is worth noting that no studies could be found that investigate the future structure of research.

5.8. Summary

This chapter established the requirements of the dataset to construct a network that was objective and takes advantage of big data availability to make it cheap, maintainable, and extendable. The University of Bath provided a suitable dataset as it allows the results to be associated with an organisation that the researcher and associates have access to. This provided an ability to double-check the results as a form of face validity.

The dataset is best described as the University of Bath co- yearly authorship data 2000-2017. This was collected from the University of Bath opus in lieu of other publicly available datasets, which are not as complete.

In addition to this, this chapter established two measures of successful that various models can be correlated to: bibliographic measures and future connectivity. Funding was discounted on the basis that no statistically significant correlations could be found. This is because the relatively few grants have been awarded to investigators with a high degree of variability, making the standard error very high. As the research aim pertains to IDR, it is necessary to develop methods to classify individuals into disciplines. This method is established in Chapter 6.

Chapter 6: Operational definition of 'disciplines'

This chapter establishes a method to classify individuals in the University of Bath yearly co-authorship dataset 2000-2017 into disciplines. Whilst disciplines are very often used in different kinds of research, it remains a construct. There is no governing body that defines what all disciplines are, nor has this been established in literature (Cobo, López-Herrera et al. 2011, Kang, Li et al. 2015).

Research often uses the term discipline interchangeably with department (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009). Research that seeks to explain why IDR is beneficial attribute the benefits to different knowledge (Guan and Liu 2016, Guan, Zuo et al. 2016). These two approaches are not equal. One is based on the classification of a person within an organisation, the other based on the knowledge classification of the person.

This chapter first establishes how this research views disciplines. It then outlines two different methods of measuring discipline: department-based disciplines and content-based disciplines. It finally reviews the validity of these methods.

6.1. Disciplines as sets

The difficulty of creating an operational definition of discipline begins with understanding the nature of disciplines. The definitions outlined in the literature review define the different type of disciplinary work that can be undertaken, but all the definitions rely on the construct of discipline.

A definition of discipline can easily be given in a dictionary:

> *"a branch of knowledge, typically one studied in higher education." – Oxford Dictionaries (2017)*

No literature disagrees with this term, but the issue is one of boundaries. If a discipline were thought of as sets, there would be a significant amount of overlap between the different sets, which makes it difficult to outright state that this person/article belongs in set A over set B if it finds itself in the overlap, as demonstrated in Figure 6.1. In the endeavour of trying to identify Persons A and B's discipline, the matter may be simple. However, Person C would either require a new category reflect $A \cap B$, or would need to be placed in some systematic way in discipline A or B. Creating a new category would draw a lot of different perspectives, whereas networks could easily show that person C is connected to both A and B. The latter is therefore preferable.

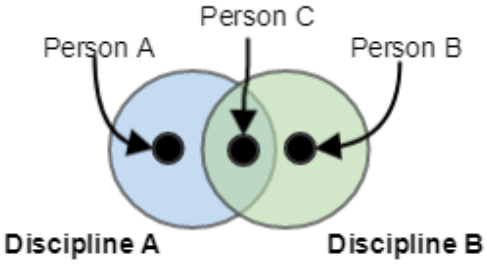

Figure 6.1. Disciplines as viewed as sets, demonstrating a significant amount of overlap.

This is particularly troublesome with academia, which seeks to further knowledge, meaning that sets need to be extrapolated and establishing where they belong and at what time is never set in stone. For instance, whilst Mechanical Engineering is well understood as a discipline, its boundaries are fuzzy, and it could in many cases be mistaken for other disciplines such as Management or Physics.

Therefore, to develop an operational definition of disciplines where an individual must belong to only one, it is necessary to develop a taxonomy of disciplines. This research proposes two different approaches.

1. Disciplines are based on organisational structure and is unique to every organisation.
2. Disciplines are based on the content of the authors' work.

6.2. Organisation-based disciplines

This method of defining disciplines was the easiest to implement as the boundaries are hard, and a person will be in one department, even if they are heavily affiliated with another. Therefore, the authors' department or centre affiliation defines their disciplines. Unfortunately, this data is not easily collected for every author in the dataset. However, an alternative approach is proposed based on the data already collected. The University of Bath publication data contained the department and centre under which the articles were published.

A simple implementation classifying individuals based on where the largest number of their publications have been published yielded decent results for the department of Mechanical Engineering. Out of 63 staff, 62 were classified in Mechanical Engineering, providing a 98.6% accuracy. This is deemed to be suitable accuracy for this research.

The same method was implemented for centres.

6.3. Content-based disciplines

The second approach to classifying authors into disciplines based on the content of their publications. This needs to be an automated process as manually classifying abstracts is subject to individual biases and is extremely resource intensive.

There are two well-established methods of analysing and classifying texts.

- Methods to group together features, such as clustering algorithms could equally be used to create groups (Arthur and Vassilvitskii 2007, Von Luxburg 2007, Aggarwal and Zhai 2012). However, this assumes that disciplines are readily separable.
- Methods to predict classification based on similarities to a training-set. The classification process could be done through various approaches. However, text data classification is amenable to machine learning techniques, especially with the data available (Smola and Schölkopf 2004, Geurts, Ernst et al. 2006). Three different phases are needed to classify every University of Bath publication into the predefined disciplines: training, testing, and predictions.

6.3.1. Concept classification

The first approach seeks to group together various concepts or keywords. To do so, three aspects need to be considered.

1. Concepts or keywords need to be extracted.
2. Some system establishing proximity or relevance of these concepts to one another.
3. A method to group the concepts into groups.

6.3.1.1. Concept extraction

There have been many different approaches to automatically extracting concepts or keywords. The majority are sophisticated and involved processes (Kaur and Gupta 2010, Parameswaran, Garcia-Molina et al. 2010, Metke-Jimenez and Karimi 2015).

One of the most commonly used unsupervised algorithms to extracting concepts or keywords are based on TextRank, a co-occurrence graph-based extractive algorithm (Mihalcea and Tarau 2004, Barrios, López et al. 2016, Allahyari, Pouriyeh et al. 2017). TextRank is directly analogous to PageRank. It uses Part-of-Speech (POS) approaches to identify concepts. Where these concepts co-occur in the same sentence, links can be established, which can then be used to find the TextRank score.

The Natural Language ToolKit (NLTK), a platform for building Python programs to process text (Bird, Loper et al. , Bird, Klein et al. 2009) provides an implementation for Rapid Automatic Keyword Extraction (RAKE). The method is similar to the TextRank but does not require POS analysis. It instead establishes concepts based on N-grams split by 'stop words' and punctuation (Rose, Engel et al. 2010). Individual words' co-occurrence strength/frequency provide the score; for higher order N-grams, the score is just the sum of its members' scores (Rose, Engel et al. 2010).

It should be noted that both these approaches do not use more advanced techniques such as synonym detection, spelling error, or alternative spellings.

Both approaches were used in this research to explore whether concept-based classification was possible.

6.3.1.2. Proximity

Having established concepts and their scores in every abstract, it was necessary to group together concepts into a discipline. A network of concepts was the most straightforward approach. The top three concepts are taken from every abstract and defined as being connected by virtue of them being in the same abstract. Only the top three are considered as every noun being included would create a network of concepts that is far too connected and create false connections between abstracts.

6.3.1.3. Identify groups

As a network of concepts was created, a network community algorithm would be able to separate concepts into groups. Whilst many algorithms would be suitable, the Louvain Modularity algorithm is suitable to dealing with large networks efficiently and effectively (Blondel, Guillaume et al. 2008). The algorithm is designed to maximise the modularity, *Q*, defined as the proportion of links inside communities compared to between communities and is given in the following expression.

$$Q = \frac{1}{2m}\sum_{ij}^{N}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{6.1}$$

Where Q is the modularity, m is the sum of all link's weights, and $\delta(c_i, c_j)$ is the Kronecker delta function, which is equal to one if the community of node i, c_i, is the same as the community of node j, c_j. The Louvain Modularity algorithm is split into two iterative steps from the outset that there are as many communities as there are nodes. The first step is to calculate the potential modularity gain of placing node *i* in the same community as its neighbour *j*. Once all potential modularity gains are found, node *i* is placed in the community that would yield the largest modularity increase. The

second phase is to construct a new network with the communities being the new nodes, and the process is repeated until the modularity is no longer increased. Finally, there is a tuning parameter that alter the size of the communities.

6.3.1.4. Findings

The produced network of concepts is shown below in Figure 6.2. The node sizes are linearly proportional to their degrees. The node colours represent different communities/partitions, which should define individual fields. 3,368 different disciplines were identified, and the 45 largest partitions only comprised of 34.82% of the concepts. The largest discipline was made of 0.87% of all concepts. Many communities were individual abstracts shown in the peripheries of the network. Clearly, better resolution is necessary, as in many cases, this suggests that an abstract is its own field.

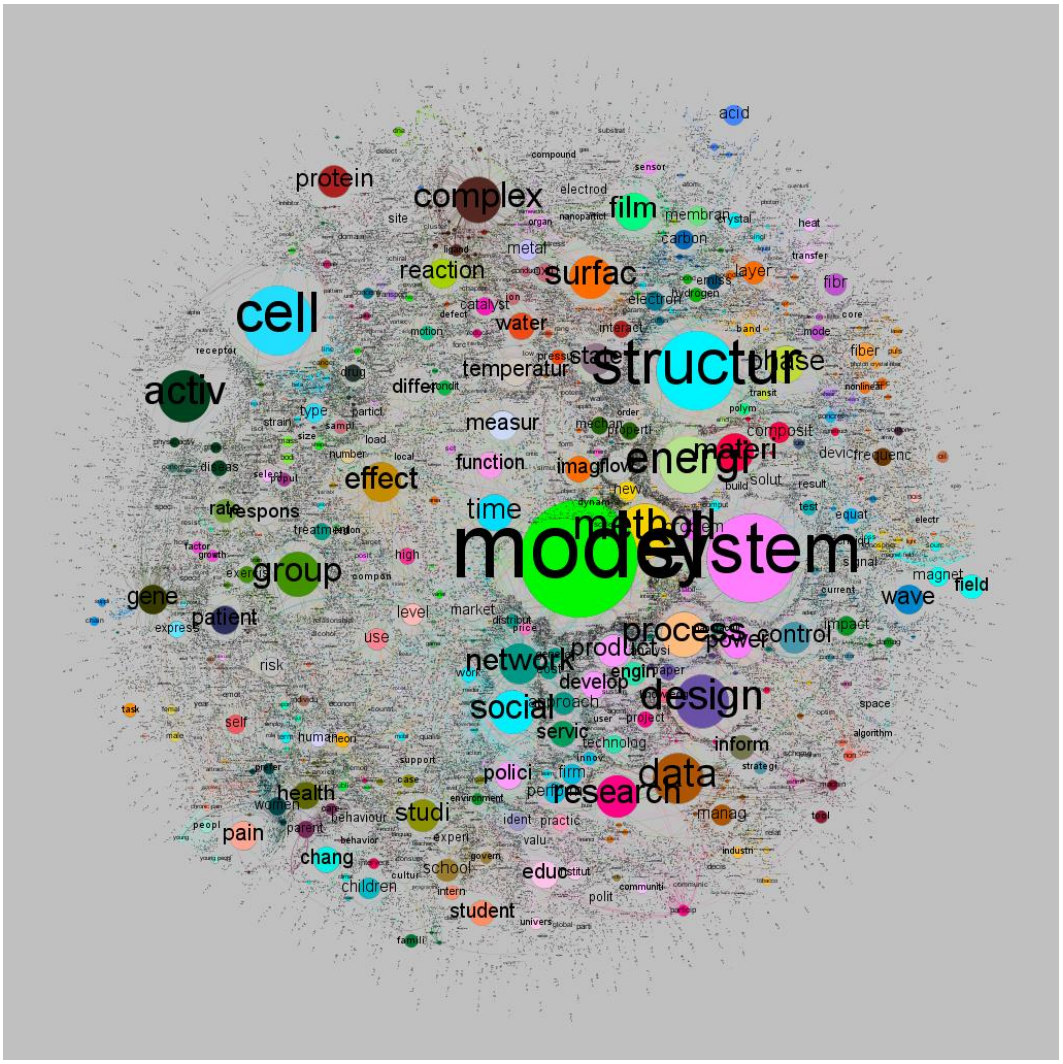

Figure 6.2. The University of Bath publications' network of concepts, 2000-2017. The concepts are formed from communities of words. These communities are detected using the Louvain algorithm.

Figure 6.3 shows a tuned Louvain algorithm (Blondel, Guillaume et al. 2008) with its resolution parameter set to 2. The node sizes are linearly proportional to their degrees. The node colours represent different communities/partitions, which should define individual fields. 2,562 different disciplines were identified, the largest containing 23.67% of all concepts, the second largest only contains 1.53%, and the third largest 0.98%.

Another problem arises: most communities join a single large cluster. Therefore, most papers would be classified under that cluster.

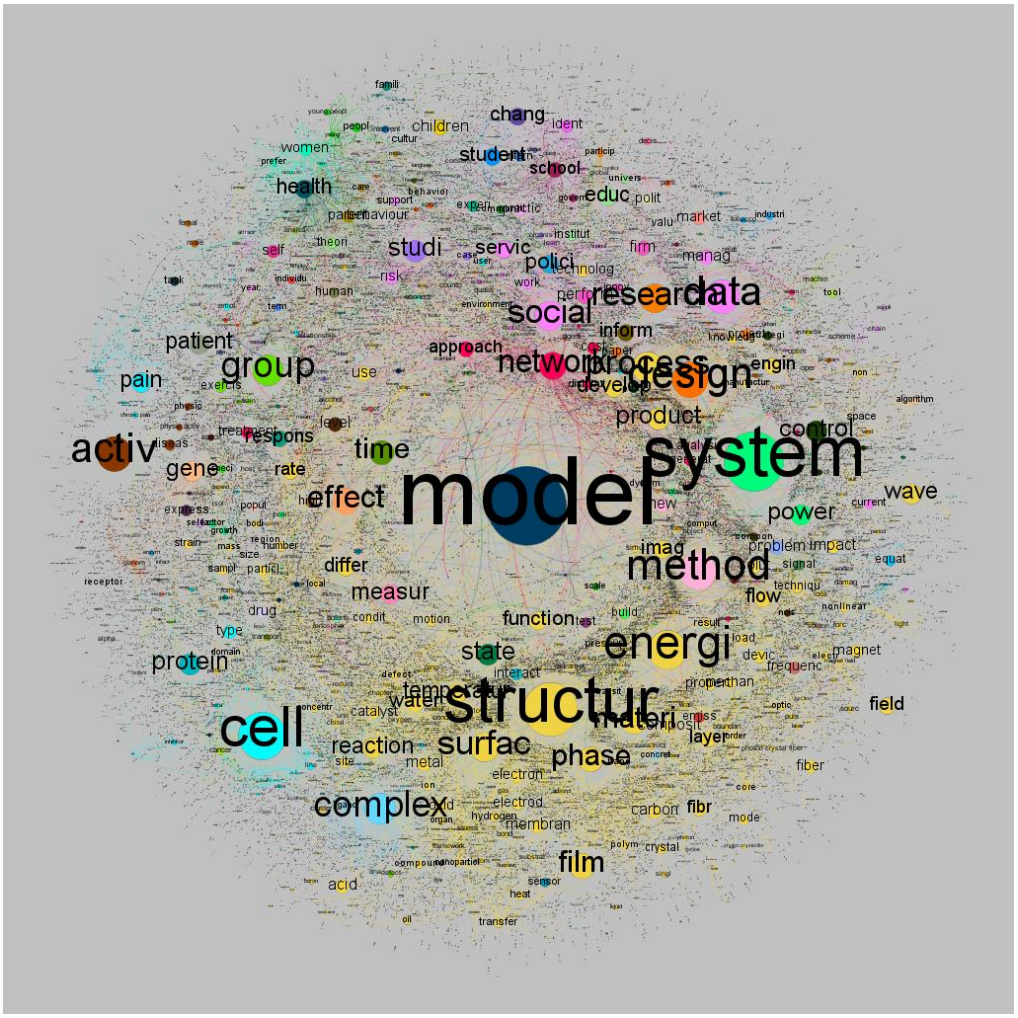

Figure 6.3. The University of Bath publications' network of concepts, 2000-2017. The concepts are formed from communities of words. These communities are detected using the Louvain algorithm. In comparison to Figure 6.2. this figure shows the communities using a tuning parameter of 2, making the communities larger. However, a giant community forms (yellow).

Furthermore, as this method does not classify abstracts into any identifiable discipline, but instead groups together concepts, it would be extremely difficult to validate the method. This is because the groups remain abstract and would have to be verified individually as there is no reference.

For these reasons, this approach is not a suitable candidate to determine discipline based on the contents. Instead an approach based on pre-determined disciplines that can be validated is necessary, making Machine Learning classification an ideal approach.

6.3.2. Machine learning classification

A second approach to determining what disciplines are is based on manually creating a taxonomy of disciplines and then classifying abstracts into these. Machine learning classification requires a program to be trained to recognise features and classify them to specific labels. It can be trained by taking cases where the features and labels are known. In the case of machine learning classification of texts, the features would be the text itself, and the labels would be the discipline it belongs in. Thus, there are three aspects to consider:

I. **Extract features:** To base the discipline based on an author's publication, it is necessary to process text from the publications. The abstract data collected is suitable for these purposes.
II. **Define labels:** Having read an abstract, the next part of the process is to classify the abstract into a discipline.
III. **Classification:** Finally, it is necessary to 'teach' an algorithm to classify the abstract into one of the disciplines.

However, to undertake each one of these, a training dataset needs to be procured. This requires sample abstracts to be collected with known disciplines. To have known disciplines, there must be a list of pre-defined disciplines.

6.3.2.1. Training dataset - Pre-determined disciplines

As there is no governing body determining commonly accepted disciplines and no review papers could be found on the matter, any created list will be subjective.

A Pre-determined list of disciplines would either need to be sourced by some relevant body qualified to define these or a bespoke list would need to be created. Creating a bespoke list needs to be created using a method of classification based on similarities and organised hierarchically. The only relevant bodies that could be found are research organisations' departments and faculties,

and professional organisations working within certain disciplines. However, each organisation is biased by its functions (e.g. there is no school of Medicine in the University of Bath).

The approach taken in this research instead is to gain an overview based on the wisdom of the crowd. Information taken from Wikipedia should always be tempered with a critical eye. Wikipedia provides two major ways to ensure quality control, every edit is recorded (making changes easy to revert) and every recent change or new page is patrolled (Wikipedia). If the wisdom of the community is to be trusted, it could be an invaluable tool for any text classification.

Based on face validity and in comparison to many Universities' departments and faculties the Wikipedia page on the outline of academic disciplines provides a balanced overview (Wikipedia). There is also an argument to be made for the user-authored nature of Wikipedia being the result of many authors' contributions, theoretically creating a better representation than any one author can. The page also provides a hierarchical view. Scraping the page for the fields yielded 28 disciplines and 634 fields lower in the hierarchy (based on CSS identifiers).

The resulting disciplines are given in Table 6.1 on the left.

However, whilst this is a reasonable list, there are several issues. Neither Management nor Finance is included. The level of aggregation is also questionable (e.g. Physics is not broken down into smaller parts whereas the Arts are into Visual Arts and Performing Arts).

As such, it was deemed that this list needed to be altered. However, this introduces the researcher's subjective perspective on the disciplines and therefore a bias (e.g. a conscious decision is made to unify the arts into a single heading,

The resulting list of disciplines are given in Table 6.1 on the right.

Table 6.1. List of disciplines detected using Wikipedia (left) and then chosen subjectively (right).

Wikipedia detected disciplines	**Subjectively chosen disciplines**
1. Agriculture and agricultural sciences 2. Biology 3. Chemical Engineering 4. Chemistry 5. Civil Engineering 6. Computer science 7. Earth and space sciences 8. Economics 9. Educational Technology 10. Electrical Engineering 11. Geography 12. History 13. Languages and literature 14. Law 15. Materials Science and Engineering 16. Mathematics (Applied) 17. Mathematics (Pure) 18. Mechanical Engineering 19. Medicine and health sciences 20. Performing arts 21. Philosophy 22. Physics 23. Political science 24. Psychology 25. Sociology 26. Systems Science 27. Theology 28. Visual Arts	1. Arts 2. Biology 3. Chemical engineering 4. Chemistry 5. Civil engineering 6. Computer science 7. Economics 8. Electrical engineering 9. Finance 10. Humanities 11. Law 12. Management 13. Manufacturing engineering 14. Mathematics 15. Mechanical engineering 16. Medicine 17. Physics 18. Psychology 19. Sociology 20. Structures and materials

The specific changes consist of:

- Inclusion of Management and Finance as disciplines
- Combining the Arts into a single discipline
- Combining ‘Mathematics (Pure)’ and ‘Mathematics (Applied)’ into a single discipline: ‘Mathematics’
- ‘Agriculture and agricultural sciences’ was deemed too specific as it was largely covered by the Engineering disciplines and was therefore removed
- ‘Political science’ is subsumed into ‘Sociology’.
- A new discipline ‘Humanities’ was defined as per the Wikipedia heading which consisted of ‘History’, ‘Geography’, ‘Languages and literature’, ‘Philosophy’, and ‘Theology’, these disciplines were subsumed into ‘Humanities’
- ‘Earth and space sciences’ were subsumed into ‘Physics’ (based on the Wikipedia content, which had a lot of link overlap).
- Subsuming ‘Systems science’ to ‘Management’

- ‘Manufacturing Engineering’ was lower in Wikipedia hierarchy. Based on its importance and focus in academia, it was included higher up in the hierarchy
- ‘Materials science and Engineering’ was renamed ‘Structures and materials’ as this was deemed to be more specific
- ‘Medicine and health sciences’ was renamed ‘Medicine’.

This is largely based on the list created by “wisdom of the crowd”, but because it has been altered, and subjective decisions have been made regarding the disciplines, it must be concluded that this list is entirely subjective.

Ultimately, a lot of subjectivity and potential sources of error have been included in creating the list. The reason that this list was used was because links could easily be found for all the disciplines, providing text data on each discipline. This provides a way to perform concept extraction for search terms for each discipline and their associated sub-disciplines (only Management and Finance required Wikipedia articles to be manually found).

Having defined a list of disciplines, it was necessary to create a training set of abstracts. Scopus provides a means of downloading abstracts. However, to do this, search terms need to be provided. Based on the Wikipedia pages, search terms could be generated. This was done by analysing the various Wikipedia content associated with each discipline. Each page was analysed, and key search terms were extracted from this raw text data using a method not unlike the concept extraction described in the previous section and used in Scopus to extract abstracts.

i. The Parts-Of-Speech (POS) from the stemmed words were found.
ii. N-grams of $N \leq 3$ were considered.
iii. Every N-gram had to contain a noun. The resulting unique N-grams were considered ‘concepts’.
iv. Every concept was given a score based on how often they appeared.
v. Three concepts from every page were chosen based on high score and how relevant they were considered to be by the researcher for every discipline.
vi. The resulting concepts were the key search terms.
vii. The source of data was searched in the abstract, keywords, title, and journal.
viii. This yielded 2,000 abstracts per search prioritised on number of citations. After removing the empty abstracts, a total of 268,774 abstracts had been recorded (mean per discipline: 12,799).

It is important to note that these search terms will introduce an incredible bias in the abstracts, which was dealt with by excluding all terms that occurred in more than 0.5% of all abstracts.

This then provides the data needed to conduct any machine learning classification. To classify all texts, the following methods were used.

However, given the way that the disciplines were defined and how the search terms were generated, there is a lot of subjective input, which ultimately will determine the performance of the classifier, and more importantly determine the boundaries of the discipline. Therefore, it is important to reflect on what the impact of inaccuracies are. There are two major aspects to consider.

First and foremost, the definition of the disciplines is the most fundamental aspect and will determine the boundaries of the research. This will have an effect that will reverberate throughout the results of the research. This is an issue of division and whether the division provide useful information. This perspective is perhaps best explained with a hypothetical disciplines list: STEM, and non-STEM. If these two were the only "disciplines" considered, then the analysis would be focused on the IDR between STEM and non-STEM researchers. This would be useful, and would certainly capture how people collaborate across very different ways of thinking. So how does breaking research down into 21 different fields provide any benefit? In exactly the same way as the hypothetical case, it is simply aimed at determining the difference between the various chosen disciplines. There is no need for the disciplines to be 'correct', but rather to simply provide a division in knowledge. The biggest issue therefore is that the difference between Physics and Chemistry will likely be greater than the differences between Mechanical Engineering and Electrical Engineering. The reason why this level aggregation was chosen is because it mirrors the departments-based disciplines, thereby providing an interesting comparison between the two (although if the content-based and department-based classification matched exactly, no new information would be gained). The second issue is that divisions are not captured. The overall effect of this would be that the statistical analysis gets a greater degree of noise. With a large enough sample size, this effect should be minimized.

Second, the generation and use of the search terms to construct a training dataset. The first issue was about creating the divisions, the second issue is actually implementing them. It was necessary to create a training-set with labelled abstracts so that unlabelled abstracts can be classified. The consequences of not accurately classifying the abstracts is that the divisions will be random and therefore not useful. By diversifying the search terms and by using many abstracts, a greater number of abstracts can be correctly classified. This was the purpose of generating search-terms. This would ultimately affect the accuracy of the predicted classifications. However, this does not change the way that any of the validation occurs. That is to say, inaccurate search terms would make the prediction inaccurate exactly in the same way that a poorly trained classifier would provide inaccurate search terms. The validation method does not change, but the accuracy may. Therefore, so long as a reasonable accuracy can be achieved, the search terms will be useful.

6.3.2.2. Procedure

The procedure followed is a standard approach that can be described in three separate phases: the training phase, testing phase, and prediction phase. This can be visualised in Figure 6.4.

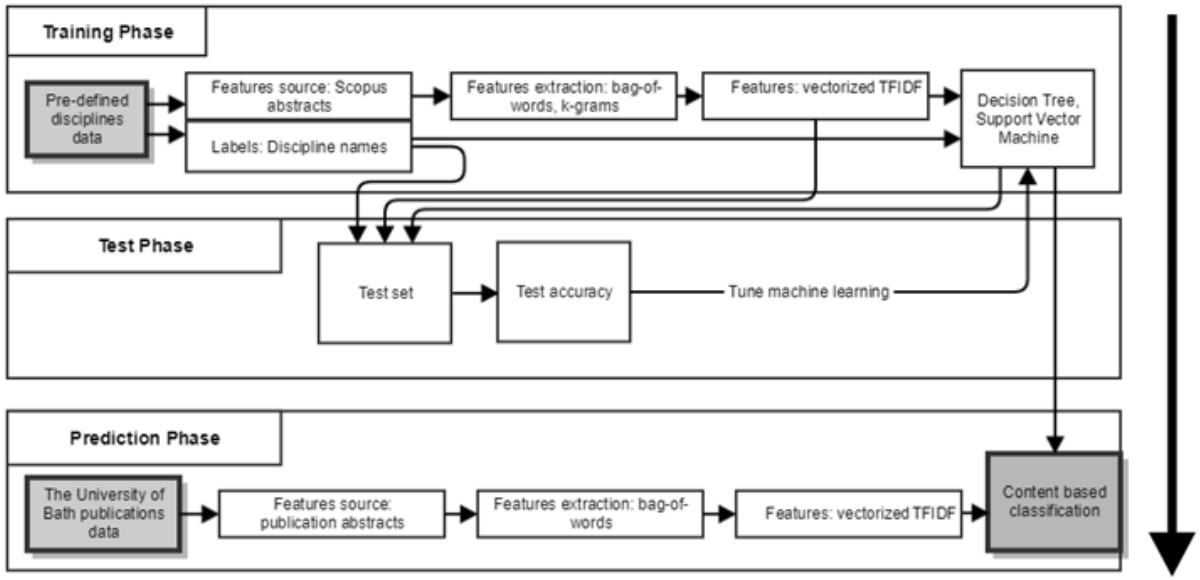

Figure 6.4. Machine learning process.

The training phase creates the classifier. The testing phase provides data on how well the program performs, and acts as validation once the program is tuned. The prediction phase follows the same process, but on the University of Bath abstracts that need classification.

6.3.2.3. Training phase

The training phase consists of many different components. In short, the training features and labels were based on the 268,774 Scopus abstracts that were collected. How these features are extracted, how they are collated, and how they are trained, had a huge impact on the accuracy of the classifier.

Feature extraction

Features can be extracted from texts in many different ways. For instance, the concept extraction above is a form of feature extraction. This was considered an approach, but ultimately a bag-of-words and N-grams approaches are better documented, and do not rely on a limited number of words from an abstract.

The text feature extraction procedure follows the guidelines laid out by the official text provided by the NLTK (Bird, Loper et al. , Bird, Klein et al. 2009). As the Scopus abstracts were recorded as strings, it was easy to use the NLTK platform to analyse individual abstracts.

The following procedure was performed.

i. The string was split-up into a list of sentences, all other punctuation was removed.
ii. The sentences were then tokenized; split into a list of words, creating a list of lists.
iii. Stopwords as defined by the NLTK platform for the English language were removed.
iv. The words were 'stemmed', using the 'Snowball' implementation (Porter).
    a. Stemming is the process of transforming a word to its roots. This therefore reduces variations of the same word to a singular form, which should be unique to that word.
    b. The Snowball method was developed by Martin Porter (Porter) as an improvement of his Porter stemmer, which utilises a combination prefix and suffix removal, reducing words on a consistent set of rules, and uses a dictionary to handle exceptions (e.g. universe and university) (Porter).
v. The text was then transformed back into a string, with full-stops to separate sentences.

The prepared text can then be used to extract measurable features. Two features types were tested. The first approach was a simple bag-of-words approach (Ko 2012, McTear, Callejas et al. 2016). This simply collects the stemmed words as individual features. This method is naïve. It does not consider the order of words, which causes problems (e.g. 'Chicago Bulls' are a famous Basketball team in the USA, a naïve method would consider 'Chicago' and 'Bulls' separately, which is unlikely to be desirable). The second approach was a k-grams approach, which uses all combinations of k adjacent words together as a single feature. Its intent is to also consider the order of words. This method considers the order of words. K-grams of 2 and 3 were considered in this research.

Having defined what features are possible, it is necessary to define what the features for a given abstract are. This research considered simple implementations of 'Term Frequency' (TF) and 'Term Frequency, Inverse Document Frequency' (TFIDF). These respectively find the number of occurrences in the abstracts and the number of occurrences in the abstract normalised by the number of occurrences across all abstracts. These methods required a large amount of data to mitigate their weaknesses.

- Synonyms are not considered in this work and every equivalent word with different roots needed to be trained for individually.
- The method is susceptible to spelling errors.

- Whilst the order of words is taken into consideration with N-grams, language structure is not. The machine learning classifier may become overly sensitive to certain adjectives which are commonly used in one field but would get false positives if those adjectives appear in other disciplines. The use of POS analyses could provide more objective features to train to.
- The University of Bath is a UK based university and will likely have a higher proportion of UK English compared to random sources. For instance, 'aeroplane' and 'airplane' will both need to be trained, but the majority of papers will likely use 'airplane' whereas the University of Bath may favour 'aeroplane'.

Label definition

The label definition was defined just as the disciplines under which the specific abstracts were searched under.

Machine learning algorithm

There are many different algorithms suitable to classify abstracts. Supervised algorithms are particularly suitable. Two different algorithms were tested as they are very well documented: a Decision Tree (DT) classifier, and a Support Vector Machine (SVM) classifier (Smola and Schölkopf 2004, Geurts, Ernst et al. 2006).

6.3.2.4. Test phase

The test phase consists of splitting the training data into training and test sets. The basic logic behind this is that a certain proportion of the training data is not actually used to create a classifier but is rather used to see how accurate the classifier is. The caveat being that it reduces the amount of training data available. It would be possible to segment training and test sets so that all data is tested/trained upon in several different runs, although this is not necessary as a significant training dataset has already been acquired.

The split chosen here is that the training set will consist of 80% of the original training set (215,019 abstracts) and the test set will consist of the other 20% of the original training set (53,755 abstracts).

If the accuracy is below what is required, it is necessary to tune the training phase either in the features that are being used, or to tune the machine learning algorithm's parameters (e.g. if the training accuracy is much higher than the testing accuracy, then the model is overfit and measures need to be taken to rectify this through various means such as Principal Component Analysis in the feature selections (Tipping and Bishop 1999), or reducing the minimum split in Decision Trees). It was in this tuning phase that a good enough accuracy was achieved by using the following parameters.

- Using n-grams approaches over bag-of-words.
- Using a TF vectorization.
- Discounting all features that appear in 0.5% of all abstracts in order to remove search-term biases.
- Adopting a Decision Tree classifier over SVM.
- Setting the minimum number of splits to 300 to avoid over-fitting the data.

All of these improved the accuracy of the classifier by improving over-fitted and under-fitted parameters.

The resulting confusion matrix is displayed below in Figure 6.5 and normalised in Figure 6.6. True negatives would have been expected in fields with similar terms (e.g. Mechanical Engineering and other engineering fields). Whilst this does occur to a small extent, it is to relatively small extent per field. Some examples do exhibit this, such as the proximity of Finance to Management, and Computer Science to Mathematics. Similar trends occur in false positives indicating that there is a fuzzy boundary between the two instead of a bias to one particular field.

However, bar a few pairings, the mistakes are evenly spread out. This could be due to two reasons.

1. The abstracts are well fitted to each other, but when occasional mistakes occur, and very different abstracts occur, the algorithm struggles.
2. The abstracts are over-fitted to the search-terms. The abstracts that contain these search-terms are well trained, whereas abstracts that do not contain these search terms cause an even spread of false positives and true negatives.

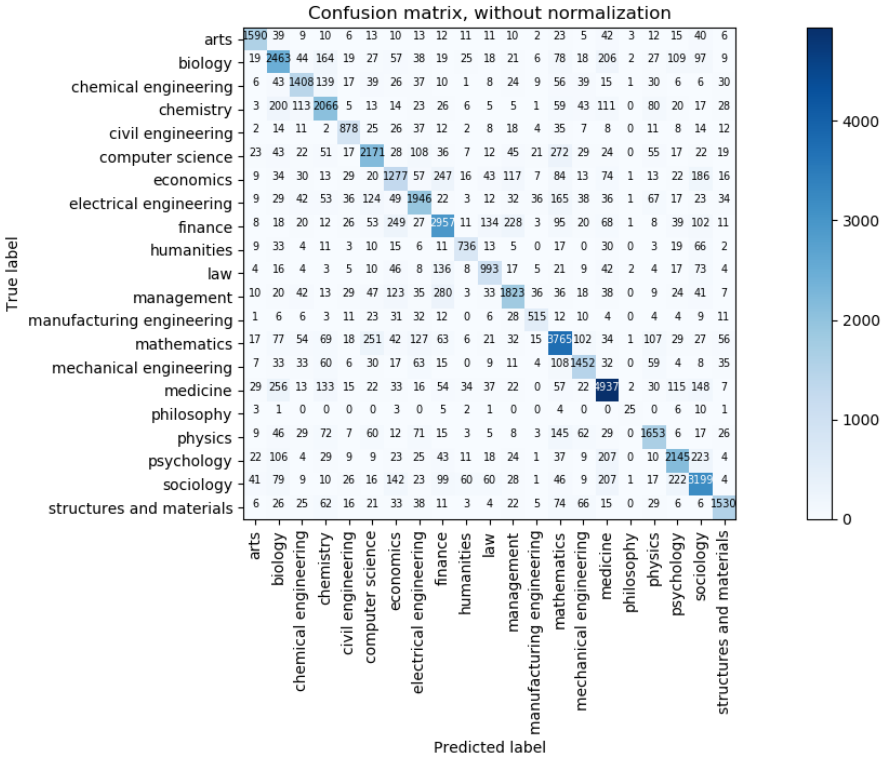

Figure 6.5. The Confusion matrix for the test phase.

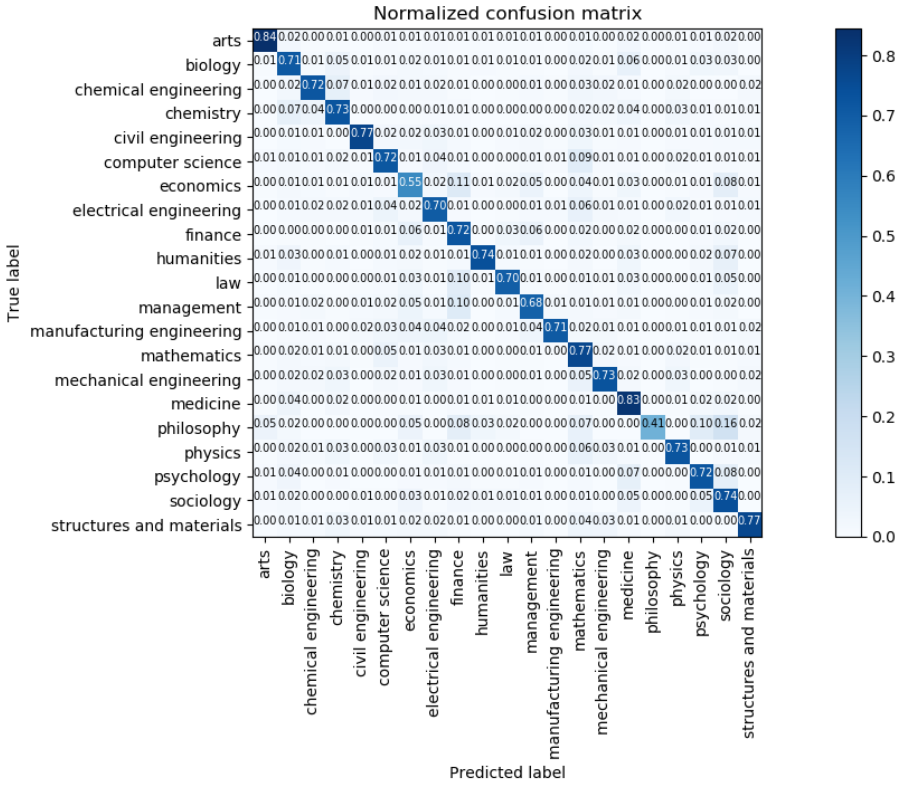

Figure 6.6. The normalised Confusion matrix for the test phase.

If there is a strong bias, the classifier would rely on the search-terms being viable to the University of Bath and the data source being large enough to gather additional features.

The test phase serves as validation for the classifier, but it is important to remember that by virtue of the data coming from the same source in the test phase, a bias is introduced against different sources. The prediction phase data came from all the University of Bath publications only, whereas the training/test data comes from the search terms. It is therefore likely that the University of Bath publications are far broader in their spectrum and the same accuracy levels would not be expected. This must be taken into consideration when considering the validation of the classifier.

6.3.2.5. Prediction phase

This phase consists of using the classifier to predict what field the abstracts from the University of Bath belong in. The same process as in the test phase, with the same parameters and training features are used.

However, whilst analysing the responses, it was noticed that the program struggled with a few terms such as chemical formulas. As a large proportion of the papers at the University of Bath came from the Chemistry department, a dictionary bias was implemented. This was done by forcing abstracts that contained key words (e.g. bromide for Chemistry, red-shift for Physics) to be associated their respective disciplines.

At this stage a complete set of disciplines have been output, but it is necessary to validate this further as the test phase may have biases that may significantly affect the accuracy. There is no way to objectively determine this as classifying abstracts would also vary from person to person.

Two forms of validation were undertaken (see Figure 6.7). The researcher evaluated the performance of 400 abstracts to ensure that there is a good fit. However, as the researcher had a bias, a further 100 abstracts were given out as surveys to 10 different individuals with at least a Bachelor's degree. The survey consisted of 10 abstracts randomly selected by the computer and a box to fill out on a scale of 1-4 (very bad – very good).

The following results were achieved.

Figure 6.7 Classifier performance. Left – The researcher's evaluation of the classifier where researcher bias may occur for the abstracts reviewed; N = 400.. Right – Survey-based evaluation (10 different individuals) as determined by resulting in N = 100.

The responses varied a significant amount from person to person, where some agreed far more with their abstract classifications, and other severely disagreeing. Ultimately, both achieved similar results with 73.75% and 69% abstracts being classified good or very good respectively.

It is likely that this number is reporting similar numbers to the test phase because it accepts approximate answers instead of exact classifications as per the test phase, which isn't necessarily wrong. For instance, if Mechanical Engineering has false positives in similar fields (for instance engineering fields and mathematics) were classified as good, the accuracy would have been 83%.

6.4. Probability requirements

It is clear that accuracies of 69-74% is not good enough, as it would mean that 26-31% of all papers are badly classified. However, it is important to remember that it is not papers that are being classified in this research, but the authors of the papers. The same method of classifying individuals as the organisation-based disciplines is proposed. Individuals are placed in the discipline that the largest number of their papers are published in.

By classifying over 50% of the papers correctly means that a large sample of papers from the same author increases the accuracy of correctly classifying individuals. This is because the probability increases with the number of publications classified into the same category. The probability of k successful trials occurring in n trials is given by the following expression.

$$P(k) = C_k^n \cdot p^k \cdot (1-p)^{n-k} \tag{6.2}$$

Where $P(k)$ is the probability of *k* successful classifications and p is the probability of successfully classifying a single paper.

$$C_k^n = \frac{n!}{k!\,(n-k)!} \tag{6.3}$$

The probability of correctly classifying an author's discipline is given in Table 6.2 (using the reported accuracy in the test phase of classifying into one of the 21 defined disciplines, 73.75%).

This means that for publications it is possible to achieve very high probabilities to classify individuals by classifying just over half of the overall number of papers per person. This would achieve about 95% accuracy in classifying a person correctly. The number of papers that need to be classified into the largest discipline is given by the following expression, provided that the accuracy is 73.35% per abstract (note this is not an analytical expression, but it holds for up to 14 publications, and is conservative beyond that point).

$$k_{n\geq 2} = \left\lceil \frac{n+1}{2} \right\rceil \tag{6.4}$$

This analysis has thus far worked on the premise that individuals fit into a specific discipline. Individuals who work in multi, cross, inter, and trans disciplines would be expected to have their abstracts correctly classified into multiple disciplines (i.e. true positives will not fit into the same group). No reliable method could be created to detect these, nor would it be a good use of resources to do so. It is assumed that individuals who fit such categories will be detected through network means.

Table 6.2 Probability of classifying people correctly with multiple papers being classified identically.

Probability for individual papers:
0.7335

Matches \ Trials	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	73.35%	53.80%	39.46%	28.95%	21.23%	15.57%	11.42%	8.38%	6.15%	4.51%	3.31%	2.43%	1.78%	1.30%
2		92.90%	82.48%	71.02%	59.80%	49.52%	40.48%	32.73%	26.24%	20.89%	16.52%	13.00%	10.18%	7.94%
3			98.11%	93.94%	87.83%	80.36%	72.14%	63.71%	55.45%	47.67%	40.53%	34.13%	28.50%	23.62%
4				99.50%	98.02%	95.30%	91.32%	86.21%	80.21%	73.61%	66.70%	59.73%	52.90%	46.40%
5					99.87%	99.37%	98.29%	96.43%	93.71%	90.11%	85.71%	80.65%	75.07%	69.16%
6						99.96%	99.81%	99.40%	98.61%	97.30%	95.39%	92.81%	89.57%	85.70%
7							99.99%	99.94%	99.80%	99.48%	98.90%	97.96%	96.59%	94.72%
8								100.00%	99.98%	99.93%	99.81%	99.57%	99.14%	98.46%
9									100.00%	99.99%	99.98%	99.93%	99.84%	99.65%
10										100.00%	100.00%	99.99%	99.98%	99.94%
11											100.00%	100.00%	100.00%	99.99%
12												100.00%	100.00%	100.00%
13													100.00%	100.00%
14														100.00%

6.5. Summary

This chapter has outlined how it is that disciplines can be measured. Two different operational definitions were proposed: department-based disciplines and content-based disciplines.

The first approach is robust and accurate.

The second approach is not as accurate, requiring more than half a person's publications to be classified into the same discipline to achieve good accuracy.

As such, the department-based disciplines are considered for the most part in this research, with the content-based disciplines used for comparison.

Chapter 7: Using networks to identify differences between disciplinary and interdisciplinary authors

This chapter outlines the correlational study conducted on the broad, historical longitudinal data. It adapts and compares several networks models that have been established in analogous studies and are positively correlated with good research outputs.

Five models were identified as being consistently present in literature and were given different reasons for their behaviour.

These are adapted by hypothesising that people who conduct IDR behave differently than individuals who work in disciplinary research. It was theorised that by identifying differences in their behaviour, a model could be created that identifies individuals who enable and sustain IDR. As such, both disciplinary and interdisciplinary authors are correlated to research outputs.

This chapter outlines how this study was conducted, presents the results, and discusses the overall implications and findings for this study. This chapter is organised as follows. The 'Approach' section outlines how the study conducted can achieve the research aim, and outlines a chapter hypothesis. The 'Methodology' section outlines the methods used to adapt the models and conduct statistical tests. Each model is then considered individually, presenting the results, the overall findings, and implications for that model. The 'Discussion' section reviews the results as a whole and discusses the implication, outlines which models hold, what their effects on IDR are, and whether the chapter hypothesis holds. The 'Conclusion' section discusses the drawbacks of the model and the needed further work to achieve the research aim.

Through this process, the study identifies that the assumption that individuals can be classified as either disciplinary or interdisciplinary does not hold, prompting a re-examination of the SNA paradigm.

7.1. Approach

This study is a correlational study that seeks to identify if established models can be adapted to identifying differences between disciplinary and interdisciplinary authors. The purpose of identifying the differences is to guide the development of a model that can identify individuals who enable and sustain IDR.

As per the research approach defined in Chapter 2, this research distinguishes scientific knowledge by whether a hypothesis has been tested and corroborated or tested and failed. Therefore, to establish whether differences have been identified, the following chapter hypothesis will be answered.

Hypothesis 7: SNA models show that there are differences between disciplinary and interdisciplinary authors.

This will be done by correlating each of the five models identified in Chapter 4 to measures of research output success. The five models are given below.

1. Degree centrality
2. Betweenness centrality
3. Eigenvector centrality
4. Structural holes
5. The strength of weak ties

Three different metrics of success were considered: an author's total impact factor (a bibliographic measure - used in lieu of better, albeit more difficult to obtain, measures such as authors' H-indices for a longitudinal study with specific start and end dates – e.g. what was a researcher's H-index in the period 2005-2008, which discounts all papers and citations that occur outside that period), the author's future degree, and funding was also considered, but remained statistically insignificant through all models tested throughout the research.

This study opted to use the impact factor as the measure of interest as future degree is a predictive measure and funding was not statistically significant as it was too sparse.

This study was conducted on the University of Bath co-authorship dataset 2000-2010 to 2000-2017. This consists of a list of papers with the authors' name, the classification of the paper (department-based and content-based), the year of publication, and the Thompson-Reuters 2015 impact factor of the journal.

7.2. Methodology

This section establishes how it is that the chapter hypothesis can be answered.

First, it details the construction of the network, which are then used to define the independent variables (the models) and the dependent variables (the metrics of success).

Secondly, it states how the models can be adapted to differentiate authors as either disciplinary or interdisciplinary.

Thirdly, it outlines the statistical analyses that are necessary to understand the difference between disciplinary and interdisciplinary authors. This includes defining the hypotheses necessary to establish differences. These various hypotheses each help answer the overall chapter hypothesis.

7.2.1. Building a network

The first step to answering the chapter hypothesis is to create the University of Bath journal collaboration network.

Two separate entities were identified: authors as nodes, and co-authorship as links.

This means there are N authors with a maximum possible number of $N(N-1)$ links. It also means that any papers with n authors will have all the authors fully connected to each other thereby creating $n(n-1)$ links.

An adjacency matrix can be created by establishing the links (co-authors) between the nodes (authors). The number of times authors have co-authored with one another provides a weight matrix.

However, as the boundaries of the research are drawn around a research organisation, the rows and columns associated with non-Bath authors are removed from these matrices, with the resulting matrices being the working network adjacency matrix and weight matrix, $\boldsymbol{A}$ and $\boldsymbol{w}$ respectively.

Through this network, all well-established network measures can be created.

Having defined how to build a network to investigate IDR, it is necessary to process the data. The data was collected in the period from 2000-2017. It is important to realise that many networks are not physical. Collaboration networks certainly are not. Collaboration is a construct that occurs over a period of time, and outside of this period, it does not exist. Therefore, collaboration networks are beholden to time.

For instance, a collaboration occurs over a specified amount of time. If the network falls outside of this period, the collaboration cannot be included in the network structure. It is for this reason that microscopic timeframes are not often used; the network structure will be very sparse and often even non-existent.

Equally, if a macroscopic timeframe is adopted, there is a loss in analytical specificity. There is no Goldilocks range, where the timeframe is ‘just right’ as every different range will show something different. The various possible frames are shown in Figure 7.1.

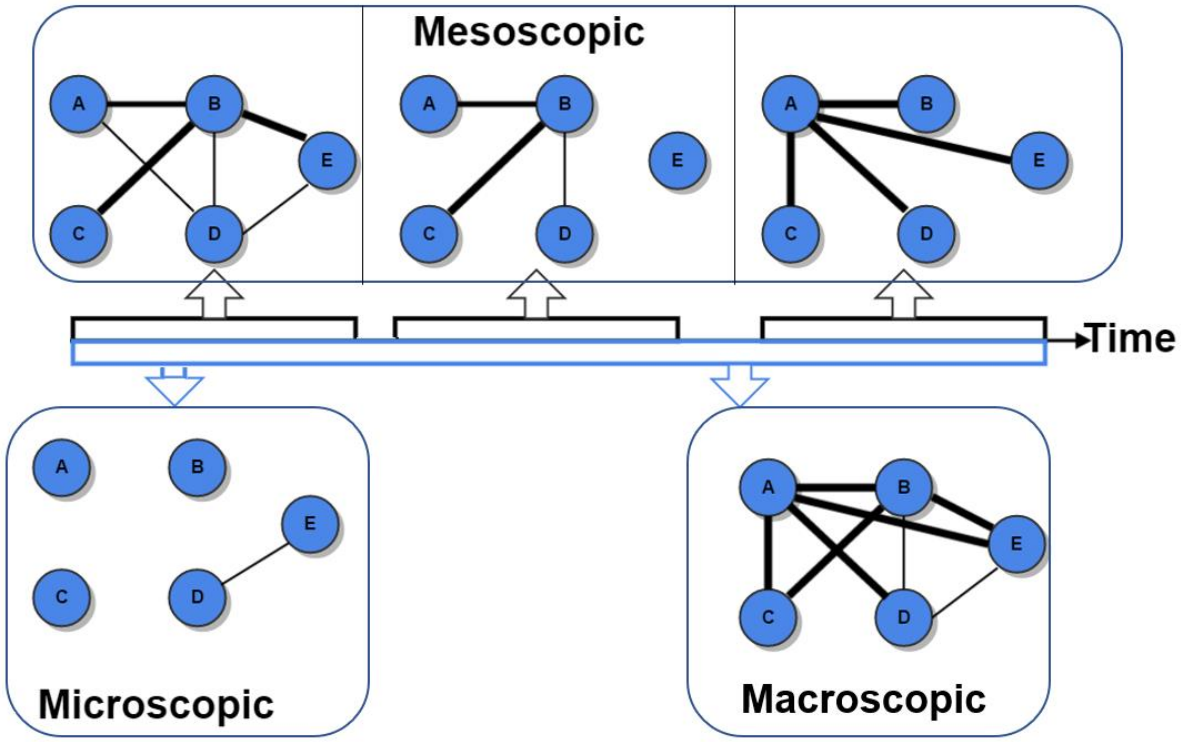

Figure 7.1. Temporal snapshots of sample networks. As time progresses from left to right, the timeframe chosen affects the network topology. Microscopic timeframes yield sparse networks. Macroscopic timeframes yield complete networks. Mesoscopic timeframes do not have a Goldilocks area, where the timeframe is just right, but rather different views that give a different amount of information.

An alternative approach is proposed where the network is analysed as a macroscopic network, but increasing the upper bound. This means that no specificity is lost, but is embedded in the yearly change.

The University of Bath co-authorship network 2000-2010 provides a non-sparse networks that is suitable as a starting point. To understand how it changes with time, the following network dates are considered: 2000-2011, 2000-2012, 2000-2013, 2000-2014, 2000-2015, 2000-2016, and 2000-2017.

This however introduces a weakness into the analysis. If a person leaves the University they would be included in the analysis whilst increasing the upper-bound. Equally, if they retain their unique Bath ID and continue to collaborate with their former colleagues, there would be no way of actually knowing that they should exist outside the boundaries of the research. This is a weakness in the research, and could potentially affect the results. However, there is no indication that this is occurring on a statistically significant level.

7.2.2. Disciplinary authors and interdisciplinary authors

Chapter 6 established what disciplines papers and individual belonged to. This study seeks to define whether individuals are disciplinary or interdisciplinary. A simple approach was taken. This study proposed that populations of IDR researchers are identified as a proportion of the research they do. If a node has x proportion of interdisciplinary collaborations, then that person will be classified as an IDR researcher. This will enable the differences between disciplinary and interdisciplinary degree to be investigated. After a few runs, this has been chosen in this research to be 0.5, setting the threshold interdisciplinary researchers at parity. This means that if at least a half of a researcher's collaborations are interdisciplinary, then they will be classified as interdisciplinary.

7.2.3. Statistical analysis of panel data

Cross-sectional analyses have been the staple approach in networks and all the measures have been focused on analysing the structure of a single cross-sectional network. However, the temporal aspect of networks has been highlighted as being of utmost importance (Holme and Saramäki 2013).

Most cross-sectional statistical analyses use the Ordinary Least Squares (OLS) methods to determine regressions and statistical significance of different populations. There are two central assumptions in such OLS methods: that every observation in the population or sample are independent, and that the dependent variable is normally distributed. So long as these assumptions hold true, simple descriptive statistics provide a powerful well-understood framework that can be used to test hypotheses.

However, normal regression methods do not hold for longitudinal analyses for three main reasons. Firstly, the assumption that each point is an independent observation does not hold. This is because longitudinal data will consist of the same individuals in different time-steps, thereby being correlated temporally (serially correlated) (McFadyen and Cannella 2004). This issue is compounded by heteroscedasticity, which affects trend goodness-of-fit. Finally, the assumption of normally-distributed distributions no longer hold (Kohler and Kreuter 2012, Buck 2015, Stock and Watson 2015).

Panel data analysis provides a method to address this. There are three main approaches to panel data: First Difference (FD) models, Fixed-Effects (FE) models, and Random-Effects (RE) models.

These models all start from the same equation format.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_t + \alpha_i + u_{it} \tag{7.1}$$

Where β_0 is the Y-intercept, β_1 is the X gradient, γ_t is the time-dependent heterogeneity (usually written as dummy variables), α_i is the unobserved heterogeneity, and u_{it} is the idiosyncratic error. The term $\eta_{it} = \alpha_i + u_{it}$ is an important quantity that represents the error in the model. This needs to be controlled for ($cov(\eta_{it}, X_{it}) = 0, \forall i, \forall t$) as it is a fundamental assumption that is necessary for panel methods to be unbiased and consistent.

The various models seek some way to deal with the error. The FD method cancels α_i by subtracting by the previous time-step. FE method cancels out by subtracting the time averaged value of Y_{it}. RE assumes that α_i is very small. The standard error of each of these methods reduces respectively. However, FE is the most robust approach. Using elements of both FE and RE are called Mixed-Effect (ME) models, and it has been argued to have the strengths of both (low standard error, and robust).

FE are very robust and well documented to control unobserved factors, which networks ultimately lack, as network measures are symptomatic of researcher's position, ability, research interest, sociability, and likely a large host of other factors that cannot be controlled (Baltagi 2008, Greene 2008). For this reason, this model is chosen.

The distribution is based on dependent variable distribution. These are all fat-tailed, and therefore require a negative binomial model to address the overdispersion.

Therefore, the following equations will be used for cross-sectional and longitudinal studies respectively:

$$\overline{Y_\iota} = \beta_0 + \beta_1 \overline{X_\iota} + \overline{\gamma_\iota} + \alpha_i + \overline{u_\iota} \tag{7.2}$$

$$Y_{it} - \overline{Y_\iota} = \beta_1 (X_{it} - \overline{X_\iota}) + \gamma_t - \overline{\gamma_\iota} + u_{it} - \overline{u_\iota} \tag{7.3}$$

Where β_1 in both cases define the trend.

7.2.4. Hypothesis testing

β_1 provides a way of defining the hypotheses. The hypotheses for each model (unless stated otherwise) is defined as follows.

Hypothesis 7.1: The model, X_{it}, is positively correlated to the metrics of success, Y_{it}.

$$Y_{it} - \overline{Y_\iota} = \beta_1 (X_{it} - \overline{X_\iota}) + \gamma_t - \overline{\gamma_\iota} + u_{it} - \overline{u_\iota}$$

Where β_0 and β_1 are constants, α_i is the unobserved heterogeneity, and $\overline{u_\iota}$ is the time-averaged idiosyncratic error. The null hypothesis and alternative hypothesis are therefore respectively defined as follows.

$$H_0: \beta_1 \leq 0$$
$$H_A: \beta_1 > 0$$

Hypothesis 7.2: Interdisciplinary authors have a different correlation to disciplinary authors.

$$(Y_{it} - \overline{Y_\iota})_{ID} - \left(\beta_{1_D}(X_{it} - \overline{X_\iota})_{ID} + (\gamma_t - \overline{\gamma_\iota} + u_{it} - \overline{u_\iota})_D\right) = \beta_1(X_{it} - \overline{X_\iota})_{ID} + \gamma_t - \overline{\gamma_\iota} + u_{it} - \overline{u_\iota}$$

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Where the sub-scripts D and ID represent disciplinary and interdisciplinary values respectively. The expression shows that the $(Y_{it} - \overline{Y_\iota})_{ID}$ is subtracted by the trend derived from disciplinary values.

It is important to note that Hypothesis 7.2 will be determined whether there is a trend in interdisciplinary nodes when the trend of the disciplinary nodes has been removed. This is called the normalised impact factor ($IF_{normalised}$). E.g. if there is a negative $IF_{normalised}$ trend, then this indicates that interdisciplinary archetypes are benefit less from having more collaborators.

This is performed for every model, using both metrics of success, and conducted separately for all authors, disciplinary authors only, and interdisciplinary authors only. This first tests whether the model holds for the University of Bath. It then tests whether there are any differences.

a) To establish the effect that all nodes' structural measures have on metrics of success. This is designed to replicate studies in similar datasets.
b) To establish the effect that disciplinary nodes' structural measures have on the metrics of success. This is to ensure that the relationship holds within a discipline and establishes a baseline to compare interdisciplinary nodes' structural measures to.
c) To establish the effect that interdisciplinary nodes' structural measures have on the metrics of success. This is the salient hypothesis for IDR. This can then be compared to disciplinary nodes' structural measures and the differences and their implications can be discussed.

7.3. Degree centrality

The degree centrality is the most well-known networks measure that has been shown to be well correlated with academic output (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009). It is therefore the logical first choice to investigate. It is a simple, and highly effective centrality that simply calculates the number of neighbours a node has. It is calculated using the following expression.

$$k_i = \sum_{j=1}^{N} \boldsymbol{A_{ij}} \tag{7.1}$$

In matrix calculations, it is simply summing each row. The strength of the measure is in its simplicity, which provides an easy to understand measure, making it far easier to draw conclusions from, or discuss its shortcomings.

Furthermore, degree centrality is widely used as it is easy to calculate with a computational cost of $O(N)$, where N is the number nodes (in this research, corresponding to 'authors'). It is also easy to understand, and to manipulate. The integer value of degrees also makes it easy to use for distributions, and histograms, which can provide very meaningful results.

This provides a topological measure which can be used to correlate to other academic factors. The basic premise behind this is that people of similar degree have similar features. That is to say that degree is indicative of the social, academic, and prominence factors, which affect how many collaborators are willing to work with that particular academic. For instance, it has been argued that degree centrality is representative of both ability and the Matthew effect (Hâncean and Perc 2016).

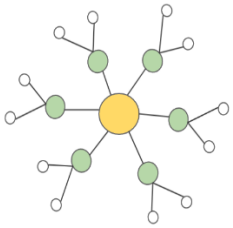

Figure 7.2. Degree centrality is depicted by the size of the node in this figure. The more connections a node has, the higher the centrality.

There is a positive correlation between degree and research output (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009), or that degree is a symptom of fitness or ability (making degree an effect of high ability) (Bianconi and Barabási 2001, Albert and

Barabási 2002). In identifying what individuals can enable and sustain IDR, this becomes invaluable.

As such, this study seeks to understand how it is that degree, and interdisciplinary degree affects metrics of output/success.

7.3.1. Model validity

The model validity is testing whether the model holds for this data set, this is done by testing Hypothesis 7.1. As can be seen in Figure 7.3, there is strong linear positive trend. The statistical results are given in Table 7.1. The F-statistic P-value is below the 0.05 threshold, and the R^2-value is 0.4270.

β_1 is 7.206 thereby rejecting the null hypothesis and accepting the alternative hypothesis. The model is validated.

This trend holds through all networks from 2000-2010 to 2000-2017.

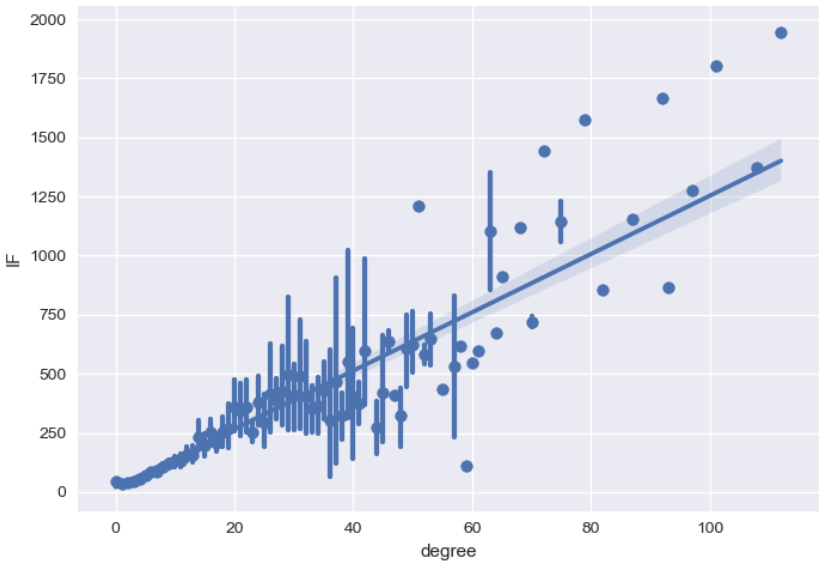

Figure 7.3. Scatter plot for all points across all time, instead of individual points being shown, bars showing the spread is shown. The clear blue band shows the 95% confidence interval for the chosen regression. The scatter plot shows the degree vs. impact factor from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A positive correlation can be seen.

Table 7.1. The statistical results of the fixed effects panel data analysis of degree vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the degree (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF	R-squared:	0.4270
Estimator:	PanelOLS	R-squared (Between):	0.3204
No. Observations:	7384	R-squared (Within):	0.5404
Date:	Mon, Jun 25 2018	R-squared (Overall):	0.3360
Time:	22:36:27	Log-likelihood	-3.462e+04
Cov. Estimator:	Clustered		
		F-statistic:	4809.7
Entities:	923	P-value	0.0000
Avg Obs:	8.0000	Distribution:	F(1,6453)
Min Obs:	8.0000		
Max Obs:	8.0000	F-statistic (robust):	82.537
		P-value	0.0000
Time periods:	8	Distribution:	F(1,6453)
Avg Obs:	923.00		
Min Obs:	923.00		
Max Obs:	923.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	46.063	3.9754	11.587	0.0000	38.270	53.856
degree	7.2057	0.7931	9.0850	0.0000	5.6509	8.7605

F-test for Poolability: 128.01
P-value: 0.0000
Distribution: F(929,6453)

7.3.2. Department-based disciplinarity differences

Having ascertained that the model remains valid overall, it is necessary to test if differences can be detected between disciplinary and interdisciplinary nodes.

The regression is shown in Figure 7.4, and a negative trend can be seen. When inspecting the statistical results in Table 7.2, the results are statistically significant, but exhibit a negative between R^2-value. This means that a horizontal fit is better for the time-averaged values. Given that an overall trend can be seen, but is not well represented by an OLS trend, no conclusion can be drawn.

However, there is a negative trend within. Upon further inspection, it can clearly be seen that the bottom interquartile range decreases year on year more than the upper interquartile range increases. The median remain time invariant. This seems to be driven by higher degree interdisciplinary authors.

Based on this information, the null hypothesis cannot adequately be rejected. Therefore, no discernible difference between disciplinary and interdisciplinary authors could be found when discerning by organisational collaborations for this model.

Table 7.2. The statistical results of the fixed effects panel data analysis of degree vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the degree (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.4206
Estimator:	PanelOLS	R-squared (Between):	-0.1416
No. Observations:	1080	R-squared (Within):	0.2985
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.0946
Time:	11:38:52	Log-likelihood	-4642.7
Cov. Estimator:	Clustered		
		F-statistic:	680.12
Entities:	917	P-value	0.0000
Avg Obs:	1.1778	Distribution:	F(1,937)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	39.142
		P-value	0.0000
Time periods:	8	Distribution:	F(1,937)
Avg Obs:	135.00		
Min Obs:	135.00		
Max Obs:	135.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	88.252	4.2869	20.586	0.0000	79.839	96.665
degree	-4.6450	0.7424	-6.2564	0.0000	-6.1021	-3.1880

F-test for Poolability: 102.28
P-value: 0.0000
Distribution: F(141,937)

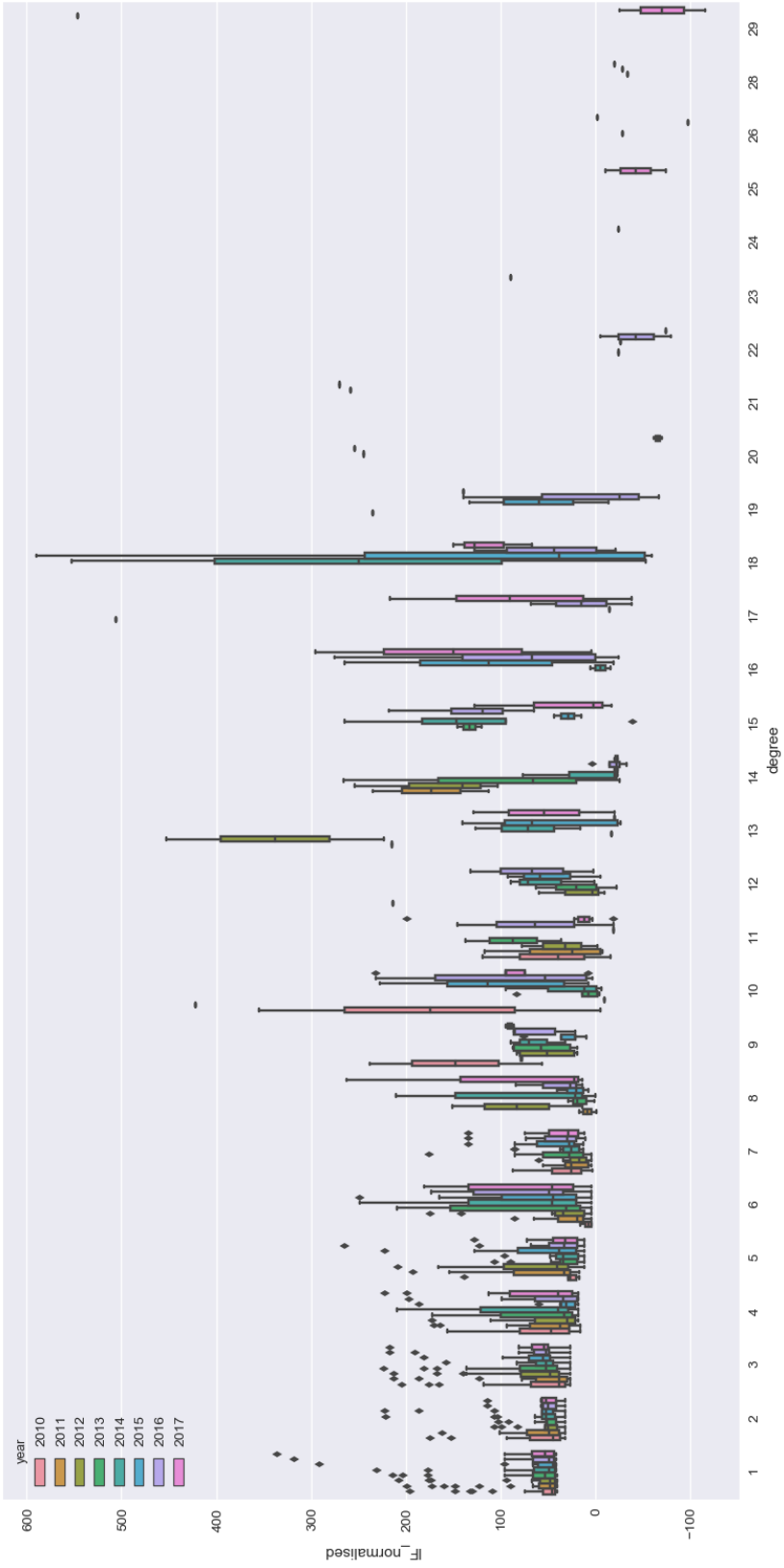

Figure 7.4. The box-plot for interdisciplinary authors as determined by department-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' degrees vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). An overall negative trend can be seen between, and inconclusive trends can be seen within.

7.3.3. Content-based disciplinarity differences

The content-based disciplinary and interdisciplinary nodes show far less variation both between and within as can be seen in Table 7.3 and Figure 7.5.

The null hypothesis cannot be rejected.

Table 7.3. The statistical results of the fixed effects panel data analysis of degree vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the degree (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	2.25e-05
Estimator:	PanelOLS	R-squared (Between):	-0.0015
No. Observations:	2832	R-squared (Within):	-0.0007
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.0015
Time:	12:26:06	Log-likelihood	-1.355e+04
Cov. Estimator:	Clustered		
		F-statistic:	0.0556
Entities:	923	P-value	0.8136
Avg Obs:	3.0683	Distribution:	F(1,2470)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	0.0008
		P-value	0.9773
Time periods:	8	Distribution:	F(1,2470)
Avg Obs:	354.00		
Min Obs:	354.00		
Max Obs:	354.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	87.027	8.0333	10.833	0.0000	71.274	102.78
degree	-0.0411	1.4435	-0.0285	0.9773	-2.8716	2.7894

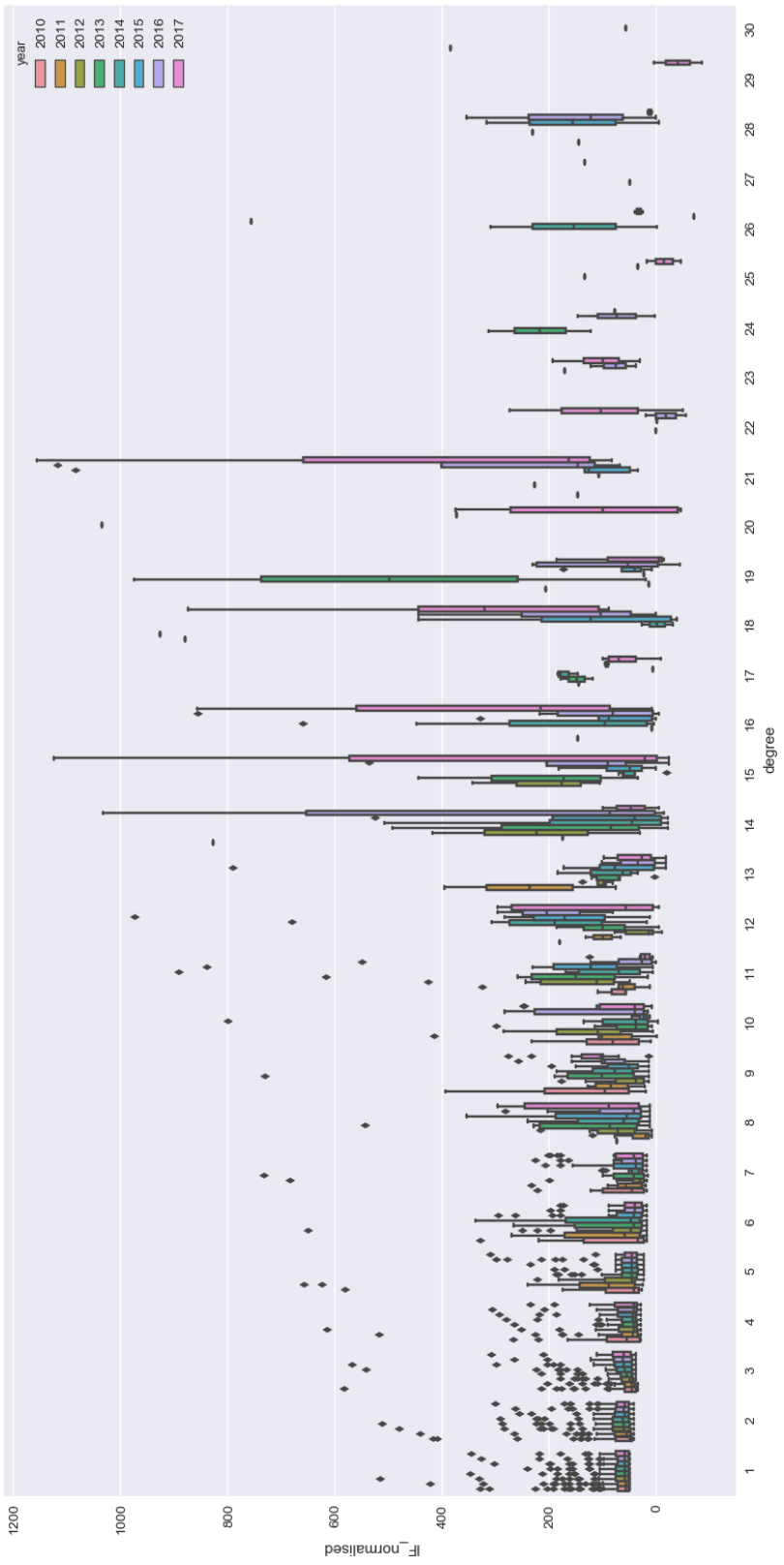

Figure 7.5. The box-plot for interdisciplinary authors as determined by content-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' degrees vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A very slight negative trend can be seen.

7.3.4. Model discussion

For the degree model, Hypothesis 7.1 is corroborated whilst Hypothesis 7.2 is rejected.

Hypothesis 7.1 being corroborated means that the findings from similar work (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009) hold for the University of Bath co-authorship network.

This suggests that the degree is a good indicator for performance in research organisations when hard boundaries are drawn around the organisation. Whilst further corroboration is needed, this could prove to be a useful metric for policy and decision makers.

However, as Hypothesis 7.2 is rejected for this model, no statistically significant differences between disciplinary and interdisciplinary authors can be found. This means that both interdisciplinary and disciplinary authors could be judged equally by their degrees.

This research aims to create a model to identify the future leaders of IDR. If the degree only highlights high degree authors, it is not specific to IDR. It equally means that it still applies to IDR, however. It can therefore be used, but not to develop an IDR specific model.

However, this was clearer for content-based disciplines as there was a clear rejection of the hypothesis. The department-based disciplines had a negative trend, but this was not statistically significant when time averaged. Therefore, there is doubt cast upon this trend. If the negative trend were true, this would imply that interdisciplinary authors stood to gain less the higher the degree is. As the constant is positive, low degree interdisciplinary nodes would actually have larger impact factor outputs than disciplinary nodes. This provides a very clear evidence-based bias towards enabling low degree nodes to conduct IDR.

7.4. Betweenness centrality

Betweenness centrality has been used to determine how in between all nodes in a network an individual is. It is calculated by determining the number of shortest paths that go through the node (Freeman 1977). Betweenness centrality is calculated by the following expression.

$$C_{betweenness_i} = \sum_{j=1, j\neq i}^{N} \sum_{k=1, k\neq\ j\neq i}^{N} \frac{\varepsilon_{jk}(i)}{\varepsilon_{jk}} \tag{7.1}$$

Where ε_{jk} is the number of shortest paths (1 if there is a definite shortest path) between nodes *j* and *k*, and $\varepsilon_{jk}(i)$ is the number of shortest paths going through node *i*. This requires two sets of information: the length of the shortest paths between pairs, and the nodes 'visited' along the path.

In cases where there is no unique path, all shortest paths between two pairs must be known. To efficiently find all paths and their dependencies, the number and length of shortest paths must be known. These can be calculated using matrix multiplication, as $\boldsymbol{A}^n$ provides the connectivity matrix for path length *n*. By increasing *n* by 1 at a time and storing the first non-zero instances in node-pairs would yield the shortest path length, *n*, and the number of shortest paths, $\boldsymbol{A}^n{}_{ij}$. However, this is computationally expensive, requiring $O(nN^3)$ calculations, where *N* is the number of nodes. Furthermore, matrix operations cannot store path dependencies.

Dijskstra and Breadth-First Search (BFS) algorithms are well suited to store the path dependencies. Furthermore, by virtue of the shortest paths acting like trees, it is possible to use predecessor paths to determine shortest paths, as shown in Figure 7.6.

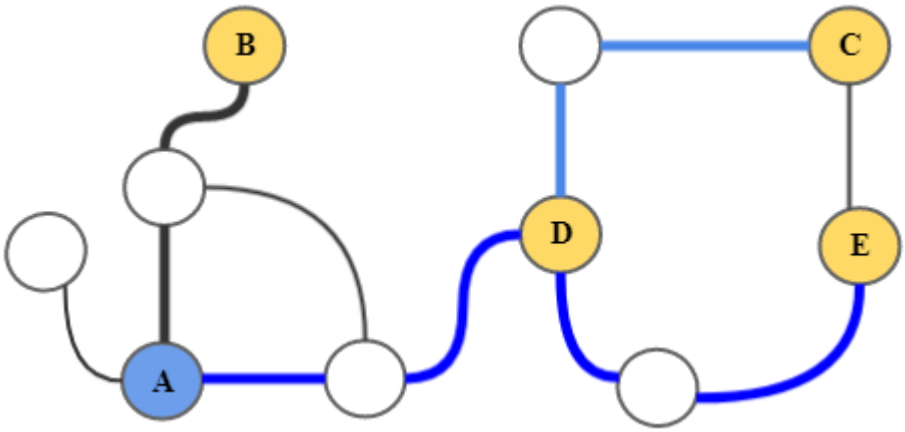

Figure 7.6. Shortest paths create trees. By utilising $path(A,C) = path(A,D) + path(D,C)$The shortest path from A to D is straightforward. The shortest path from D to C and D to E diverge. The shortest path(A, C) is equal to the shortest path(A, D) + path(D, C). In exactly the same way, this can be taken advantage of to reduce the computational cost.

Using such methods, it is also possible to reduce the computational cost of finding the shortest paths to $O(Ne + N^2 \log N)$ and $O(Ne)$ in unweighted networks, where *e* is the number of links (Brandes 2001).

Betweenness has been reasoned as being important to developing academic knowledge as it provides an indication of how many different ideas flow through a node, thereby increasing their overall centrality in the knowledge network (Nahapiet and Ghoshal 1998, Li, Liao et al. 2013). This assumes heterogeneous knowledge flows, but a positive correlation to academic outputs has been found.

7.4.1. Model validity

The model validity is testing whether the model holds for this data set, this is done by testing Hypothesis 7.1. As can be seen in Figure 7.7, there is linear positive trend. The statistical results are given in Table 7.1. The F-statistic P-value is below the 0.05 threshold, and the R^2-value is 0.2390. This is a relatively weak correlation, and does not perform as well as the degree centrality model.

β_1 is 0.0031 thereby rejecting the null hypothesis and accepting the alternative hypothesis. The model is validated.

This trend holds through all networks from 2000-2010 to 2000-2017.

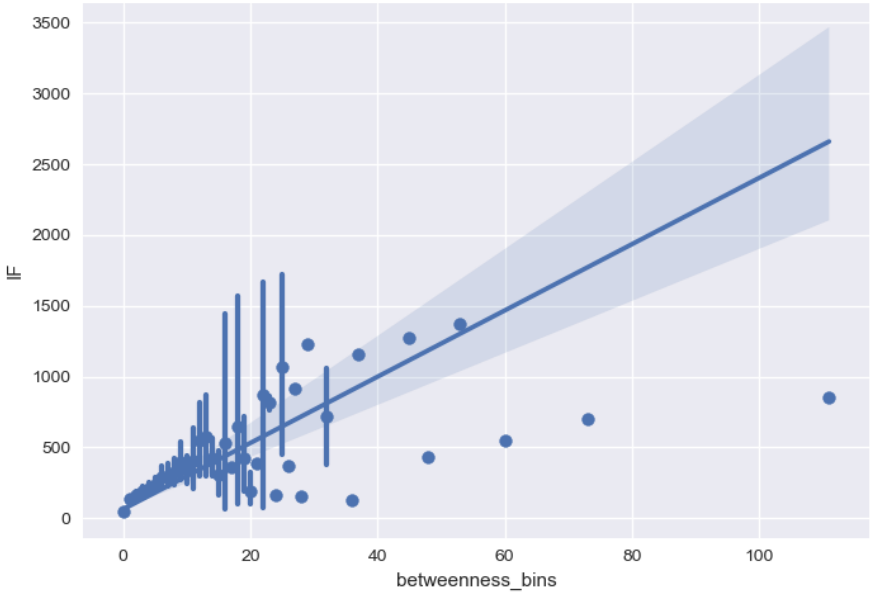

Figure 7.7. Scatter plot for all points across all time, instead of individual points being shown, bars showing the spread is shown. The clear blue band shows the 95% confidence interval for the chosen regression. The scatter plot shows the betweenness vs. impact factor from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A positive correlation can be seen.

Table 7.4. The statistical results of the fixed effects panel data analysis of betweenness vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the betweenness (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF	R-squared:	0.2390
Estimator:	PanelOLS	R-squared (Between):	0.1518
No. Observations:	7384	R-squared (Within):	0.3028
Date:	Tue, Jun 26 2018	R-squared (Overall):	0.1625
Time:	15:22:01	Log-likelihood	-3.567e+04
Cov. Estimator:	Clustered		
		F-statistic:	2026.8
Entities:	923	P-value	0.0000
Avg Obs:	8.0000	Distribution:	F(1,6453)
Min Obs:	8.0000		
Max Obs:	8.0000	F-statistic (robust):	25.145
		P-value	0.0000
Time periods:	8	Distribution:	F(1,6453)
Avg Obs:	923.00		
Min Obs:	923.00		
Max Obs:	923.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	72.792	1.8720	38.884	0.0000	69.122	76.462
betweenness	0.0031	0.0006	5.0145	0.0000	0.0019	0.0043

F-test for Poolability: 120.74
P-value: 0.0000
Distribution: F(929,6453)

7.4.2. Department-based disciplinarity differences

As the model is deemed valid, it is possible to test for differences in disciplinary and interdisciplinary authors.

The box-plot of the correlation is shown in Figure 7.8. No trend can be seen between, and a small positive increase can be seen within. The trend given in the statistical analysis confirms that it is a very small value in Table 7.5. This is not valid as the R^2-value between is negative. This means that a horizontal fit is better for the time-averaged values. Given that an overall trend can be seen, but is not well represented by an OLS trend, no conclusion can be drawn.

The null hypothesis cannot be rejected. Hypothesis 7.2 is rejected and no discernible difference between disciplinary and interdisciplinary authors can be identified for the department-based disciplines for the betweenness model.

Table 7.5. The statistical results of the fixed effects panel data analysis of betweenness vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is no trend between and a weak trend within.

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.3769
Estimator:	PanelOLS	R-squared (Between):	-0.1644
No. Observations:	1080	R-squared (Within):	0.1803
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.1248
Time:	15:44:57	Log-likelihood	-4733.0
Cov. Estimator:	Clustered		
		F-statistic:	566.84
Entities:	917	P-value	0.0000
Avg Obs:	1.1778	Distribution:	F(1,937)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	103.53
		P-value	0.0000
Time periods:	8	Distribution:	F(1,937)
Avg Obs:	135.00		
Min Obs:	135.00		
Max Obs:	135.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	130.39	1.1854	109.99	0.0000	128.06	132.71
betweenness	-0.0024	0.0002	-10.175	0.0000	-0.0028	-0.0019

F-test for Poolability: 89.487
P-value: 0.0000
Distribution: F(141,937)

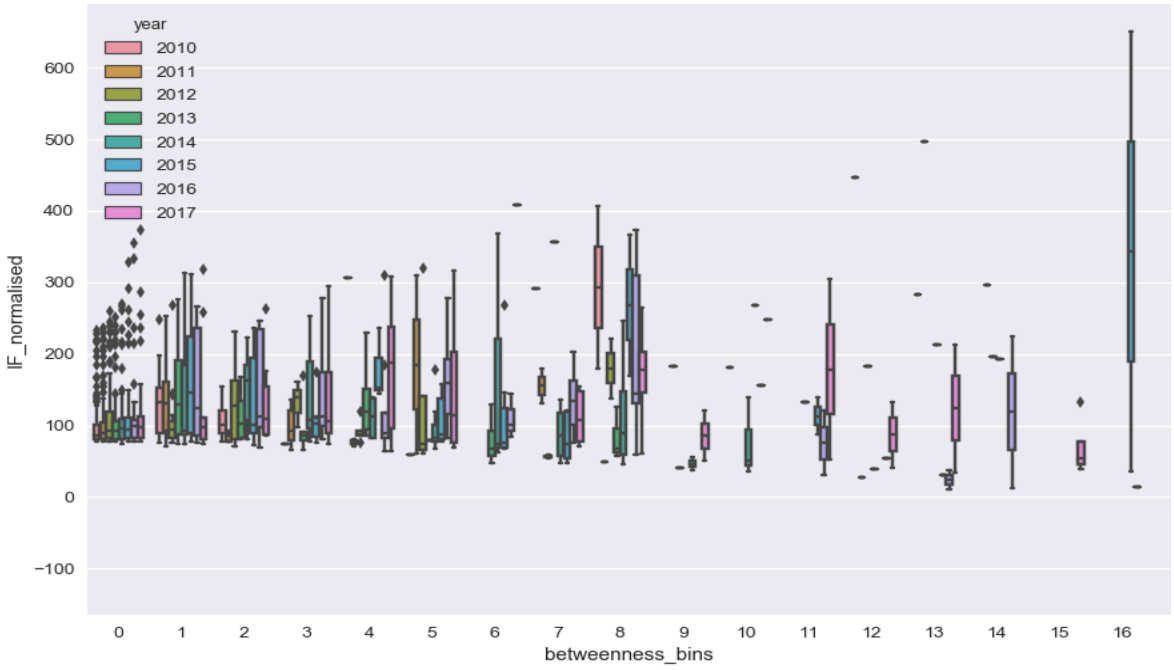

Figure 7.8. The box-plot for interdisciplinary authors as determined by department-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' betweenness vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). An overall negative trend can be seen between, and inconclusive trends can be seen within.

7.4.3. Content-based disciplinarity differences

The content-based disciplinary again shows less variation between disciplinary and interdisciplinary authors as shown in Table 7.6.

The null hypothesis cannot be rejected, and therefore Hypothesis 7.2 is rejected. There is no discernible difference between disciplinary and interdisciplinary nodes for content-based disciplines for the betweenness model.

Table 7.6. The statistical results of the fixed effects panel data analysis of betweenness vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the betweenness (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.1694
Estimator:	PanelOLS	R-squared (Between):	-0.0858
No. Observations:	2832	R-squared (Within):	0.0984
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.0694
Time:	16:48:46	Log-likelihood	-1.39e+04
Cov. Estimator:	Clustered		
		F-statistic:	503.82
Entities:	923	P-value	0.0000
Avg Obs:	3.0683	Distribution:	F(1,2470)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	13.467
		P-value	0.0002
Time periods:	8	Distribution:	F(1,2470)
Avg Obs:	354.00		
Min Obs:	354.00		
Max Obs:	354.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	139.67	2.1132	66.094	0.0000	135.52	143.81
betweenness	-0.0020	0.0005	-3.6697	0.0002	-0.0031	-0.0009

F-test for Poolability: 89.392
P-value: 0.0000
Distribution: F(360,2470)

7.4.4. Model discussion

For the betweenness model, Hypothesis 7.1 is corroborated whilst Hypothesis 7.2 is rejected.

Hypothesis 7.1 being corroborated means that the findings from similar work (Li, Liao et al. 2013) hold for the University of Bath co-authorship network.

This suggests that the betweenness is a suitable indicator for performance in research organisations when hard boundaries are drawn around the organisation. However, it has a weaker correlation to impact factor than the degree centrality does.

As with the degree centrality, Hypothesis 7.2 is rejected and no statistically significant differences between disciplinary and interdisciplinary authors can be found. This means that both interdisciplinary and disciplinary authors could be judged equally by their betweenness centrality.

As with the degree centrality, this can be useful as there is a trend, but it provides us with no further knowledge regarding the differences between disciplinary and interdisciplinary authors.

7.5. PageRank centrality

The eigenvector centrality has had mixed results in literature. Studies suggest that because Eigenvector centralities can be high by connecting to highly connected nodes, it is actually inversely proportional to academic output (Cimenler, Reeves et al. 2014).

This section tests this assertion and if there are differences between disciplinary and interdisciplinary authors.

To implement this model, a pure Eigenvector centrality is not used, as it can lack robustness in implementation. The PageRank centrality is an Eigenvector centrality and will be used in lieu as it deals with non-fully connected networks with ease, and provides high computational efficiency (Page, Brin et al. 1999).

$$C_{PR_i} = \frac{1-d}{N} + d\sum_{j=1}^{N} \boldsymbol{A_{ij}} \cdot \frac{C_{PR_j}}{k_j} \tag{7.1}$$

Where d is a damping factor, typically valued at 0.85. This needs to be calculated iteratively, until the values of every node converges. This is typically achieved in less than 100 time steps. This then applies a PageRank score for every node.

Applying this to the University of Bath co-authorship network allowed the hypotheses regarding the Eigenvector to be tested.

7.5.1. Model validity

As can be seen in Figure 7.9, there is a positive trend. The statistical results are given in Table 7.7. The F-statistic P-value is below the 0.05 threshold, and the R^2-value is 0.0510, but 0.2547 between. This is a relatively weak correlation, and does not perform as well as the degree centrality model.

β_1 is 33720 thereby confirming the null hypothesis and rejecting the alternative hypothesis. The model that high eigenvector centrality will be less successful is disproved for the University of Bath co-authorship network 2000-2010 to 2000-2017.

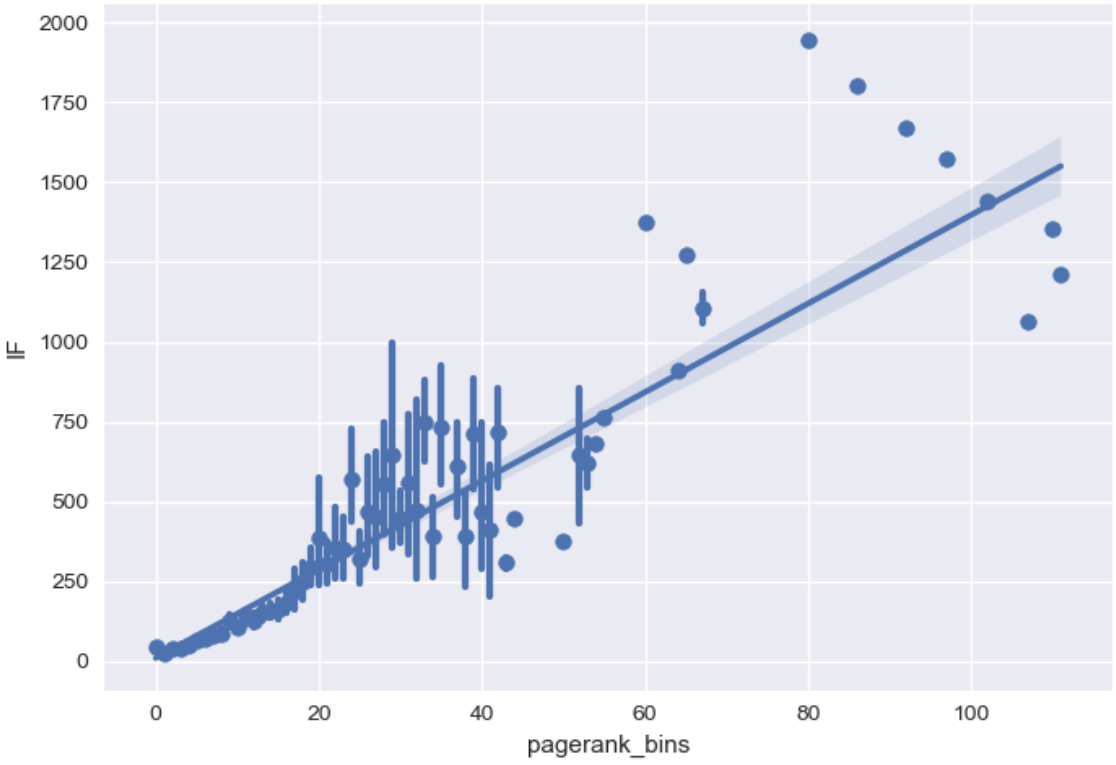

Figure 7.9. Scatter plot for all points across all time, instead of individual points being shown, bars showing the spread is shown. The clear blue band shows the 95% confidence interval for the chosen regression. The scatter plot shows the PageRank centrality vs. impact factor from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A positive correlation can be seen.

Table 7.7. The statistical results of the fixed effects panel data analysis of PageRank centrality vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the PageRank centrality (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF	R-squared:	0.0510
Estimator:	PanelOLS	R-squared (Between):	0.2547
No. Observations:	7384	R-squared (Within):	0.0021
Date:	Tue, Jun 26 2018	R-squared (Overall):	0.2368
Time:	18:38:26	Log-likelihood	-3.648e+04
Cov. Estimator:	Clustered		
		F-statistic:	346.97
Entities:	923	P-value	0.0000
Avg Obs:	8.0000	Distribution:	F(1,6453)
Min Obs:	8.0000		
Max Obs:	8.0000	F-statistic (robust):	6.1175
		P-value	0.0134
Time periods:	8	Distribution:	F(1,6453)
Avg Obs:	923.00		
Min Obs:	923.00		
Max Obs:	923.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	55.316	10.861	5.0931	0.0000	34.025	76.608
pagerank	3.372e+04	1.364e+04	2.4734	0.0134	6995.1	6.045e+04

F-test for Poolability: 68.646
P-value: 0.0000
Distribution: F(929,6453)

7.5.2. Department-based disciplinarity differences

The box-plot of the correlation is shown in Figure 7.10. The is a very weak negative trend. However, as can be seen in Table 7.8, whilst the overall trend is statistically significant, R^2-value between is negative. However, the trend within is positive.

This suggests that the correlation trend between the page rank and the impact factor could be becoming stronger for interdisciplinary authors. This is better visualised in Figure 7.11.

The null hypothesis cannot be rejected. Hypothesis 7.2 is rejected and no discernible difference between disciplinary and interdisciplinary authors can be identified for the department-based disciplines for the betweenness model.

Table 7.8. The statistical results of the fixed effects panel data analysis of betweenness vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is no trend between and a weak trend within.

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.3769
Estimator:	PanelOLS	R-squared (Between):	-0.1644
No. Observations:	1080	R-squared (Within):	0.1803
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.1248
Time:	15:44:57	Log-likelihood	-4733.0
Cov. Estimator:	Clustered		
		F-statistic:	566.84
Entities:	917	P-value	0.0000
Avg Obs:	1.1778	Distribution:	F(1,937)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	103.53
		P-value	0.0000
Time periods:	8	Distribution:	F(1,937)
Avg Obs:	135.00		
Min Obs:	135.00		
Max Obs:	135.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	130.39	1.1854	109.99	0.0000	128.06	132.71
betweenness	-0.0024	0.0002	-10.175	0.0000	-0.0028	-0.0019

F-test for Poolability: 89.487
P-value: 0.0000
Distribution: F(141,937)

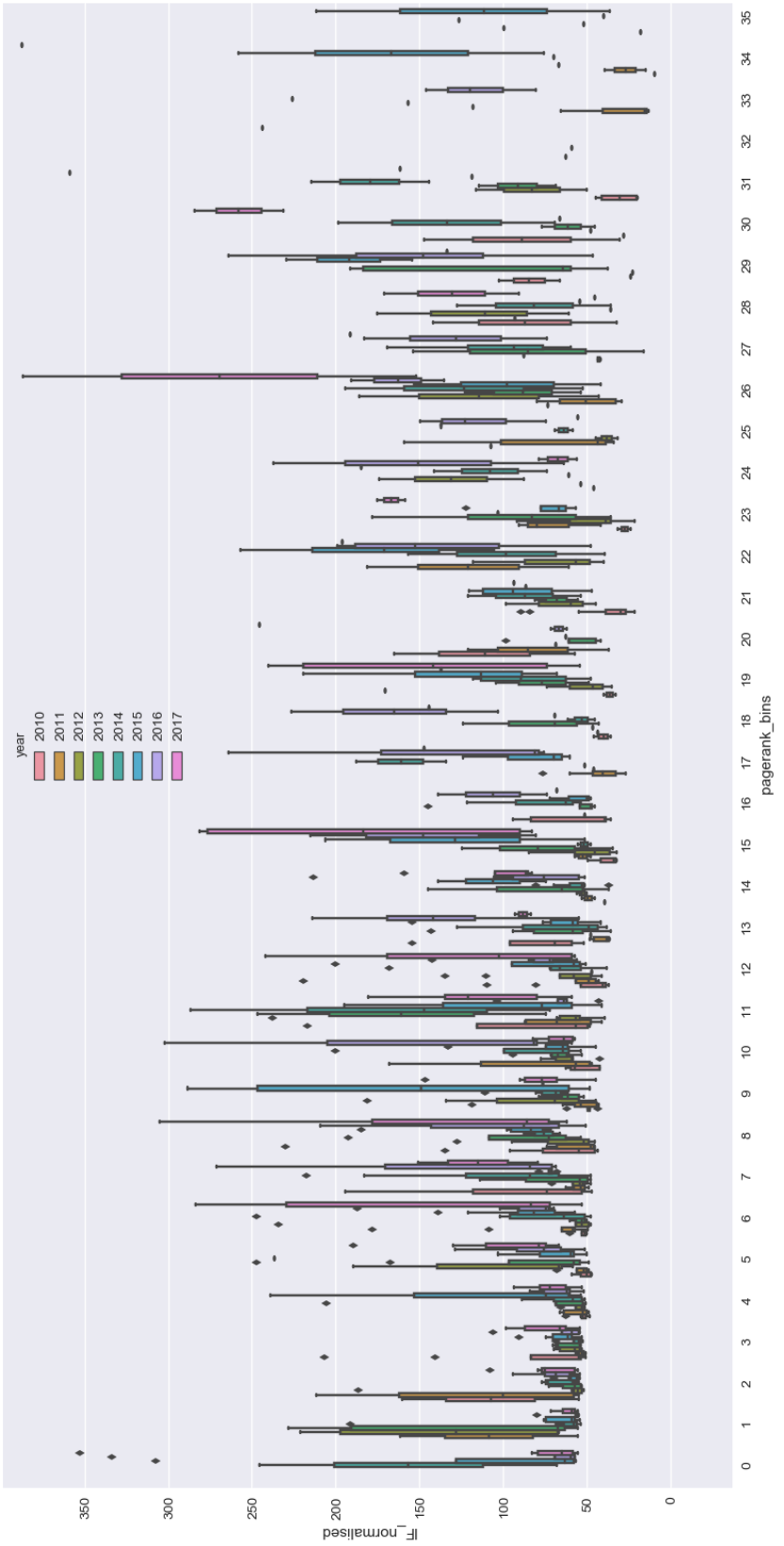

Figure 7.10. The box-plot for interdisciplinary authors as determined by department-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' PageRank centrality vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). An overall negative trend can be seen between, and inconclusive trends can be seen within.

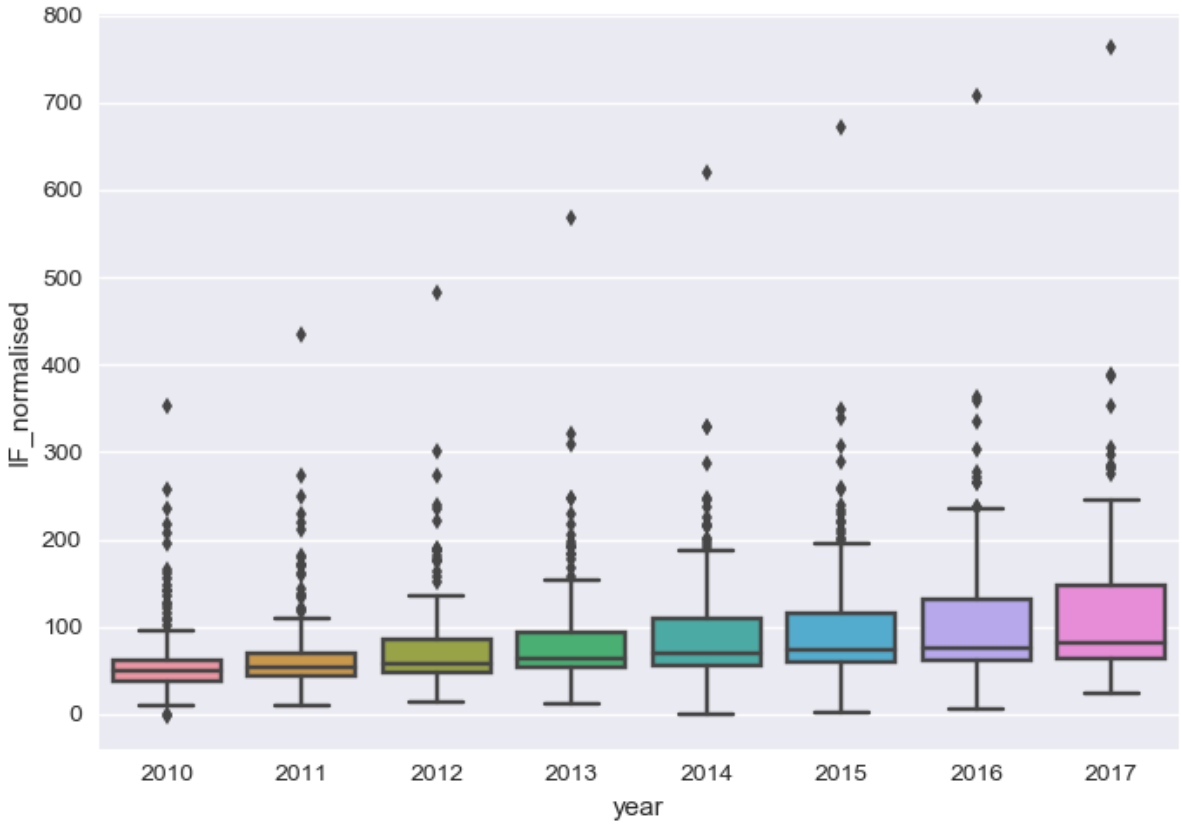

Figure 7.11. Box-plot of the impact factor normalised by the disciplinary PageRank trend.

7.5.3. Content-based disciplinarity differences

The content-based disciplines show a statistically significant difference between the interdisciplinary and disciplinary authors as can be seen in Figure 7.12 and Table 7.9. It is time-invariant as can be seen in the within R^2-value, but has a β_1 value of 60390 with an R^2-value of 0.3667 between.

The null hypothesis is rejected, and the alternative hypothesis is corroborated. Hypothesis 7.2 therefore passes.

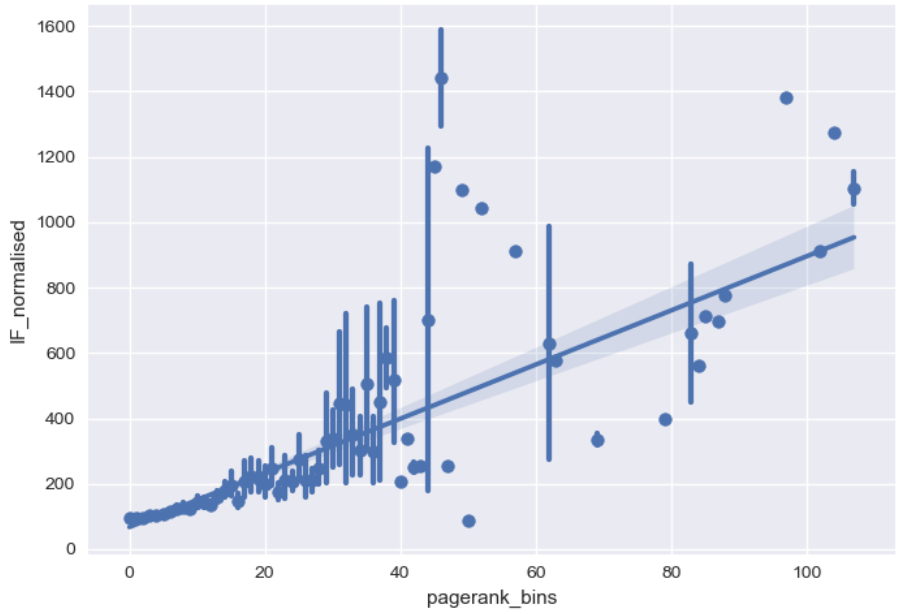

Figure 7.12. Scatter plot for all points across all time, instead of individual points being shown, bars showing the spread is shown. The clear blue band shows the 95% confidence interval for the chosen regression. The scatter plot shows the PageRank centrality vs. impact factor from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A positive correlation can be seen.

Table 7.9. The statistical results of the fixed effects panel data analysis of PageRank centrality vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the PageRank centrality (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.1481
Estimator:	PanelOLS	R-squared (Between):	0.3667
No. Observations:	2832	R-squared (Within):	-0.0065
Date:	Tue, Jun 26 2018	R-squared (Overall):	0.3316
Time:	19:09:40	Log-likelihood	-1.399e+04
Cov. Estimator:	Clustered		
		F-statistic:	429.36
Entities:	923	P-value	0.0000
Avg Obs:	3.0683	Distribution:	F(1,2470)
Min Obs:	0.0000		
Max Obs:	8.0000	F-statistic (robust):	6.3663
		P-value	0.0117
Time periods:	8	Distribution:	F(1,2470)
Avg Obs:	354.00		
Min Obs:	354.00		
Max Obs:	354.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	91.767	21.226	4.3234	0.0000	50.145	133.39
pagerank	6.039e+04	2.393e+04	2.5232	0.0117	1.346e+04	1.073e+05

F-test for Poolability: 57.624
P-value: 0.0000
Distribution: F(360,2470)

7.5.4. Model discussion

For the Eigenvector centrality/PageRank model, Hypothesis 7.1 was rejected. The null hypothesis holds. The null hypothesis is inline with the other centrality models, and it can be said that centrality measures are positively correlated to the impact factor in the University of Bath co-authorship network, which is corroborated by the degree, betweenness, and PageRank centralities.

Hypothesis 7.2 is rejected for the department-based disciplines, but is corroborated for the content-based disciplines. This marks the first identification that there is a difference between disciplinary and interdisciplinary authors.

This corroboration can provide clues as to why there is a difference, which could potentially be used to create a model.

7.6. Structural holes

Structural holes in collaboration networks has been linked with diversity of information. As such, it would be expected that IDR would be characterised by a lot of structural holes. Structural holes are defined by absence of links between a node's neighbours. For instance, neighbours to node *i, j* and *k* who are not linked to one another would be creating a structural hole.

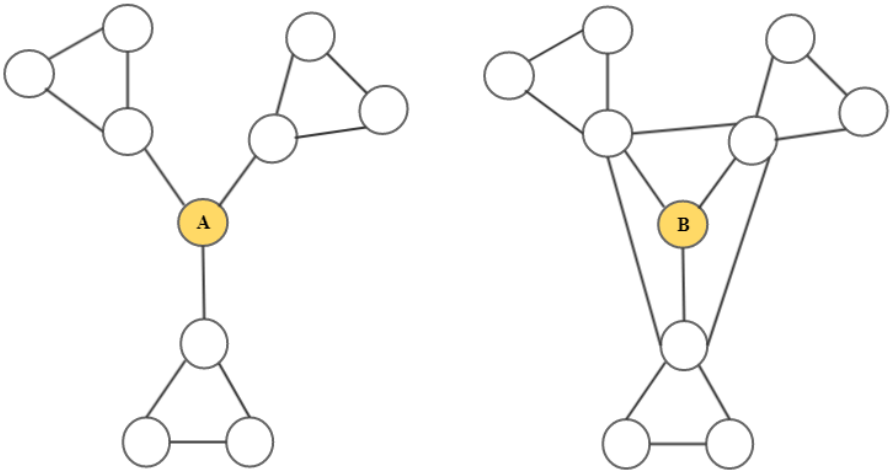

Figure 7.13. The two structures show that node A has three structural holes, whereas node B does not have any. Despite the second graph being denser, it is reasoned that node A benefits from greater diversity. There is greater redundancy in the structure on the right.

The original concept was defined the individuals having complementary knowledge, but not being directly connected (Burt 2004, Burt 2009). This concept is based on the same underlying reasoning as the strength of weak ties that suggests that stronger, more homophilious ties are more likely to overlap in neighbours. That is to say, different knowledge travels through 'bridges', which are unlikely to have redundant connections.

It is important to note that such network structures are hypothetical, and based on a small-world concept, whilst having entire communities that are only weakly connected to other communities is rare.

The number of paths of length k leading from vertex i to j can be given by ${A^k}_{i,j}$. Therefore, the number of triangles is given by the following expression.

$$number\ of\ triangles\ at\ node\ i = {A^3}_{i,i} \tag{7.1}$$

A variation was found to provide better fits when indirect closures were also included in the measure. Indirect closures not only measures whether there is a direct link between neighbours, but also if there is an indirect link between them via another node.

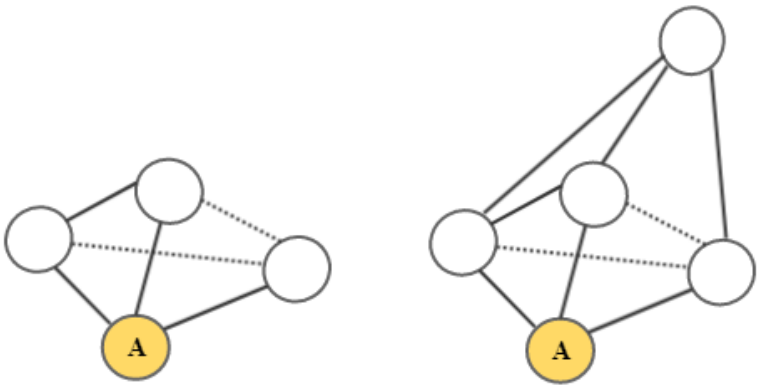

Figure 7.14. Two different structures are considered in this figure. The circles represent nodes, solid lines represent links, and the dashed lines represent structural holes affecting node A. Consider the structure shown on the left. There are two structural holes. If a node were to be added as shown on the right, it is arguable that there is an indirect closure affecting the structural holes, lessening their impact.

This therefore not only accounts for triangles, but for rectangles as well. The number of rectangles associated with a node can be calculated by considering that paths of length 4 starting and finishing at the same node consists repeat movements of the second order and squares. Thus, subtracting such paths from ${A^4}_{i,i}$ would yield the number of directional rectangles associated at a given node. Using the following equation, the number of rectangles can be found.

$$number\ of\ rectangles\ at\ nodes\ i = {A^4}_{i,i} - \left({{A^2}_{i,i}}^2 + \sum_{j=1, j\neq i}^{N} {A^2}_{i,j} \right) \tag{7.2}$$

However, this unfortunately does not consider multiple intermediate nodes connecting to the same two neighbours (i.e. if there are many different indirect paths between two neighbours). Having additional indirect path would increase the number of rectangles, and this is not desired for calculating a structural hole, which is simply interested in if there is a closure or not. Therefore, the number of rectangles forming unique indirect closures will be given by the following expression.

$$number\ of\ rectangles\ at\ nodes\ i = {A^4}_{i,i} - \left({{A^2}_{i,i}}^2 + \sum_{j=1, j\neq i}^{N} \left(\begin{cases} {A^2}_{i,j} & if {A^2}_{i,j} \leq 2 \\ 2 & if\ {A^2}_{i,j} > 2 \end{cases} \right) \right) \tag{7.3}$$

$$number\ of\ rectangles\ at\ nodes\ i = {A^4}_{i,i} - \left({{A^2}_{i,i}}^2 + \sum_{j=1, j\neq i}^{N} {A^2}_{i,j_{x\leq 2}} \right) \tag{7.4}$$

Having established how to find the number of closures, it is then important to calculate the number of possible closures given the number of neighbours. By subtracting the number of closure by this number, the number of structural holes is found. The maximum number of structural holes is given by the following expression.

$$\sigma_{max_{k\geq 2}} = \frac{k!}{(k-2)!} \tag{7.5}$$

This is derived from the unique number of permutations possible from neighbour pairs. Therefore, the proportion of possible triangular and rectangular structural holes are given by the following expressions.

$$\sigma_{i_{tri_{k_i\geq 2}}} = \frac{k!}{(k-2)!} - {A^3}_{i,i} \tag{7.6}$$

$$\sigma_{i_{quad_{k_i\geq 2}}} = \frac{k!}{(k-2)!} - \left({A^4}_{i,i} - \left({{A^2}_{i,i}}^2 + \sum_{j=1, j\neq i}^{N} {A^2}_{i,j_{x\leq 2}} \right) \right) \tag{7.7}$$

Whilst both can be combined to calculate a single number, these should be weighted. It is proposed that the structural hole contribution should be dependent on the path length. The dependency should furthermore be exponential on the path length of the closures (3 and 4 for triangular and rectangular closures respectively). Normalising by $(l-2)^2$ provides the following expressions.

$$\sigma_{i_{tri_{k_i\geq 2}}} = \frac{k!}{(k-2)!} - {A^3}_{i,i} \tag{7.8}$$

$$\sigma_{i_{quad}}{}_{k_i \geq 2} = \frac{1}{4}\left(\frac{k!}{(k-2)!} - \left(A^4{}_{i,i} - \left(A^2{}_{i,i}{}^2 + \sum_{j=1, j\neq i}^{N} A^2{}_{i,j}{}_{x\leq 2}\right)\right)\right) \tag{7.9}$$

These are both an inversion analogous to the closed triplets clustering coefficient (Wasserman and Faust 1994).

$$C = 3 \cdot \frac{number\ of\ closed\ triplets}{number\ of\ connected\ triplets} \tag{7.10}$$

Summing these two provides a single measure that can be used to measure structural holes. This measure is expected to be proportional to the metrics of success. The full measure is defined as the following expression.

$$\begin{aligned}\sigma_{i_{indirect}}{}_{k_i \geq 2} &= \frac{k!}{(k-2)!} - A^3{}_{i,i} \\ &+ \frac{1}{4}\left(\frac{k!}{(k-2)!} - \left(A^4{}_{i,i} - \left(A^2{}_{i,i}{}^2 + \sum_{j=1, j\neq i}^{N} A^2{}_{i,j}{}_{x\leq 2}\right)\right)\right)\end{aligned} \tag{7.11}$$

Using this measure, it is possible to investigate the effect of structural holes on research and IDR.

It is important note that this assumes unweighted networks. To apply weight to clustering is not straightforward. Opsahl and Panzarasa (2009) outline that various approaches to weighted clustering have been proposed. The method proposed by Barrat, Barthelemy et al. (2004) suggests a measure using the arithmetic means of triplets. This method is robust, but does not take into consideration the weight of the closure (a closure has a length of 3, a triplet has a length of 2). As no suitable weighted measure can be found, an unweighted measure is used, although this could certainly be improved upon.

7.6.1. Model validity

The model validity is based on two things. The first is that the entire premise is based on structural holes representing heterogeneous knowledge. The second is that heterogeneity is associated with better academic outputs.

Therefore Hypothesis 7.1 needs to be extended to include both.

i. The proportion of structural holes, σ_i, is greater in N_{inter} than in N_{intra}.:

$$H_0: \mu(\sigma_{inter}) \leq \mu(\sigma_{intra})$$
$$H_A: \mu(\sigma_{inter}) > \mu(\sigma_{intra})$$

ii. The model, X_{it}, is positively correlated to the metrics of success, Y_{it}.

$$Y_{it} - \overline{Y_i} = \beta_1(X_{it} - \overline{X_i}) + \gamma_t - \overline{\gamma_i} + u_{it} - \overline{u_i}$$

Part *i.* was tested on the University of Bath network 2000-2017, and is tested using a cross-sectional two-tailed t-test. The results are shown in Table 7.10. It indicates that there is no statistically significant difference between the structural holes means for disciplinary and interdisciplinary authors.

This means that the premise that interdisciplinary authors have access to more diverse knowledge is not perceptible through the network structure.

Part *ii.* is tested to see if there is a positive correlation between the structural holes measure and the impact factor. As can be seen in Figure 7.15, there is a positive trend. The statistical results are given in Table 7.11. The F-statistic P-value is below the 0.05 threshold, and the R^2-value is 0.4768.

The null hypothesis is rejected and Hypothesis 7.1 *ii.* is corroborated. This implies that the number of structural holes a node has is positively correlated to the impact factor.

Table 7.10. T-test statistical analysis of structural holes measure.

The mean department-based disciplinary structural holes measure, $\langle \sigma_{intra} \rangle$	9.3461
The mean department-based interdisciplinary structural holes measure, $\langle \sigma_{inter} \rangle$	11.4954
T-value:	-1.5854
P-value:	0.11322
The mean content-based disciplinary structural holes measure, $\langle \sigma_{intra} \rangle$	9.1620
The mean content-based interdisciplinary structural holes measure, $\langle \sigma_{inter} \rangle$	10.4617
T-value:	-1.3186
P-value:	0.18762

β_1 is 33720 thereby confirming the null hypothesis and rejecting the alternative hypothesis. The model that high eigenvector centrality will be less successful is disproved for the University of Bath co-authorship network 2000-2010 to 2000-2017.

However, as there is a positive trend, it is proposed that the null hypothesis model replaces the model.

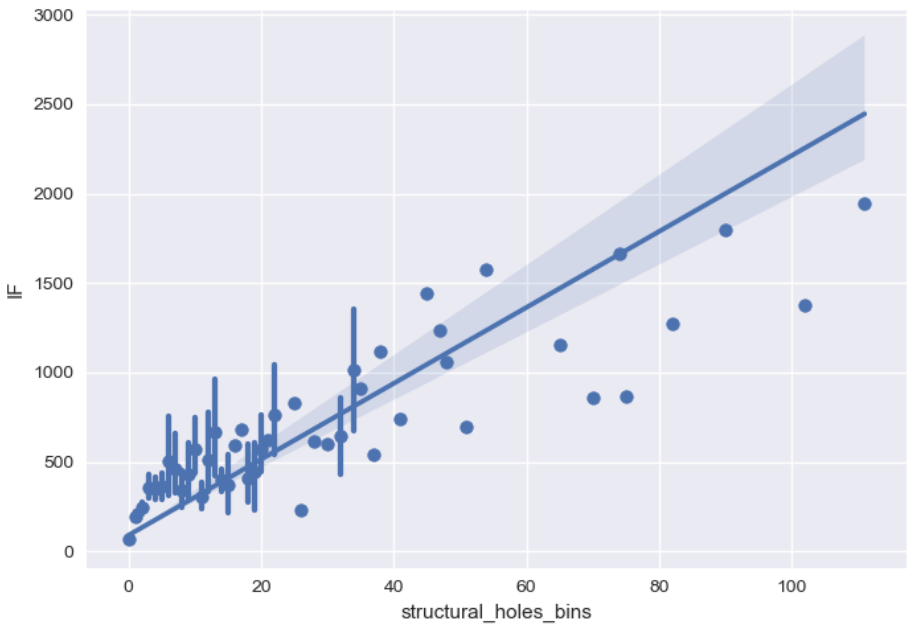

Figure 7.15. Scatter plot for all points across all time, instead of individual points being shown, bars showing the spread is shown. The clear blue band shows the 95% confidence interval for the chosen regression. The scatter plot shows the structural holes vs. impact factor from 2000-2010 to 2000-2017 (i.e. 8 time-periods). A positive correlation can be seen.

Table 7.11. The statistical results of the fixed effects panel data analysis of structural holes vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the structural holes (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF	R-squared:	0.4768
Estimator:	PanelOLS	R-squared (Between):	0.1992
No. Observations:	4655	R-squared (Within):	0.4899
Date:	Tue, Jun 26 2018	R-squared (Overall):	0.2290
Time:	22:20:20	Log-likelihood	-2.205e+04
Cov. Estimator:	Clustered		
		F-statistic:	3594.5
Entities:	703	P-value	0.0000
Avg Obs:	6.6216	Distribution:	F(1,3944)
Min Obs:	1.0000		
Max Obs:	8.0000	F-statistic (robust):	77.440
		P-value	0.0000
Time periods:	8	Distribution:	F(1,3944)
Avg Obs:	581.88		
Min Obs:	432.00		
Max Obs:	703.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	96.727	1.1859	81.564	0.0000	94.402	99.052
structural_holes	0.0998	0.0113	8.8000	0.0000	0.0776	0.1220

F-test for Poolability: 155.28
P-value: 0.0000
Distribution: F(709,3944)

7.6.2. Department-based disciplinarity differences

The box-plot of the correlation is shown in Figure 7.16. There appears to be a positive trend. However, this is not statistically significant as can be seen in Table 7.12.

If the trend were statistically significant, it would imply that interdisciplinary authors stand to benefit more from structural holes.

However, as it is not statistically significant, the null hypothesis cannot be rejected, and Hypothesis 7.2 is not corroborated.

Table 7.12. The statistical results of the fixed effects panel data analysis of betweenness vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is no trend between and a weak trend within.

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	1.292e-07
Estimator:	PanelOLS	R-squared (Between):	-9.799e-05
No. Observations:	896	R-squared (Within):	0.0001
Date:	Tue, Jun 26 2018	R-squared (Overall):	6.109e-05
Time:	22:34:39	Log-likelihood	-3832.7
Cov. Estimator:	Clustered		
		F-statistic:	9.873e-05
Entities:	124	P-value	0.9921
Avg Obs:	7.2258	Distribution:	F(1,764)
Min Obs:	1.0000		
Max Obs:	8.0000	F-statistic (robust):	3.735e-06
		P-value	0.9985
Time periods:	8	Distribution:	F(1,764)
Avg Obs:	112.00		
Min Obs:	91.000		
Max Obs:	124.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	166.67	1.8175	91.700	0.0000	163.10	170.24
structural_holes	6.216e-05	0.0322	0.0019	0.9985	-0.0631	0.0632

F-test for Poolability: 113.13
P-value: 0.0000
Distribution: F(130,764)

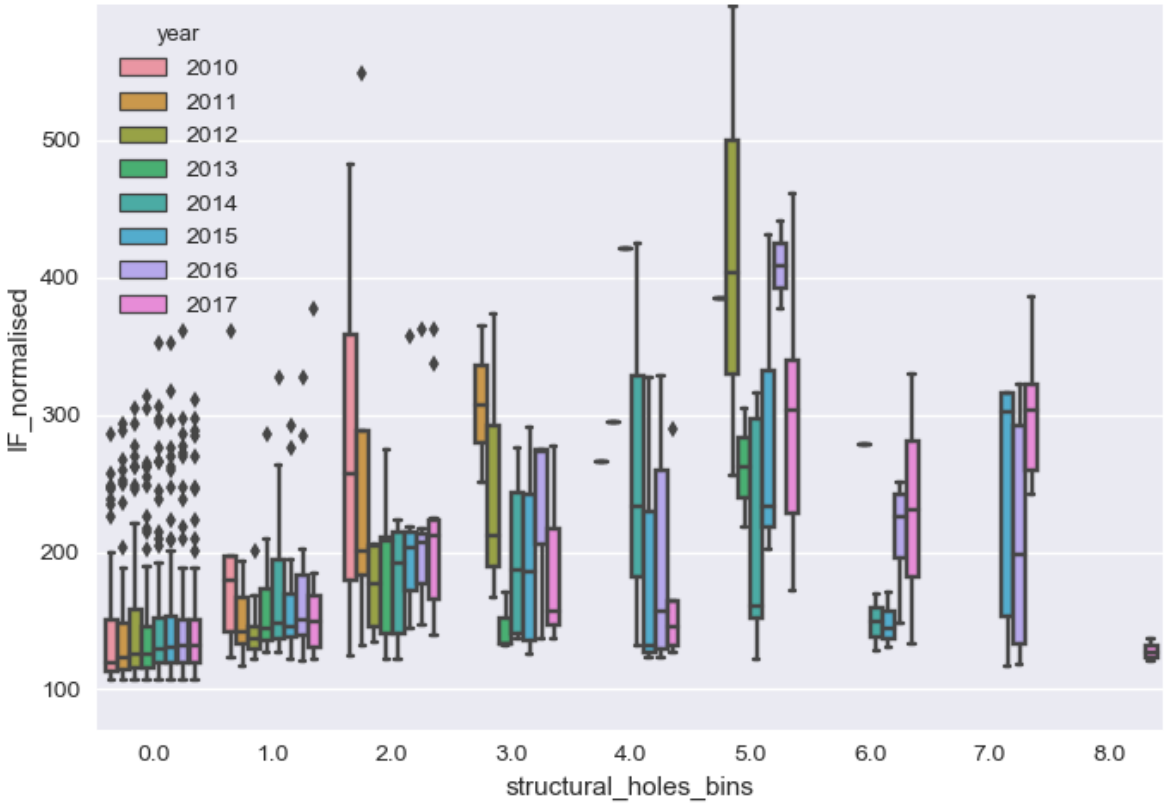

Figure 7.16. The box-plot for interdisciplinary authors as determined by department-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' structural holes vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). An overall negative trend can be seen between, and inconclusive trends can be seen within.

7.6.3. Content-based disciplinarity differences

The content-based disciplines show similar results, albeit with more randomness. Hypothesis 7.2 is therefore not corroborated.

Table 7.13. The statistical results of the fixed effects panel data analysis of structural holes vs impact factor from 2000-2010 to 2000-2017. The R-squared values show that there is a relatively strong positive trend based on the structural holes (between) and along time (within).

PanelOLS Estimation Summary

Dep. Variable:	IF_normalised	R-squared:	0.0107
Estimator:	PanelOLS	R-squared (Between):	-0.0194
No. Observations:	2171	R-squared (Within):	-0.0100
Date:	Tue, Jun 26 2018	R-squared (Overall):	-0.0188
Time:	22:40:07	Log-likelihood	-1.041e+04
Cov. Estimator:	Clustered		
		F-statistic:	20.069
Entities:	302	P-value	0.0000
Avg Obs:	7.1887	Distribution:	F(1,1861)
Min Obs:	1.0000		
Max Obs:	8.0000	F-statistic (robust):	1.0264
		P-value	0.3111
Time periods:	8	Distribution:	F(1,1861)
Avg Obs:	271.38		
Min Obs:	219.00		
Max Obs:	302.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	191.17	0.9895	193.19	0.0000	189.22	193.11
structural_holes	-0.0118	0.0116	-1.0131	0.3111	-0.0347	0.0110

F-test for Poolability: 118.82
P-value: 0.0000
Distribution: F(308,1861)

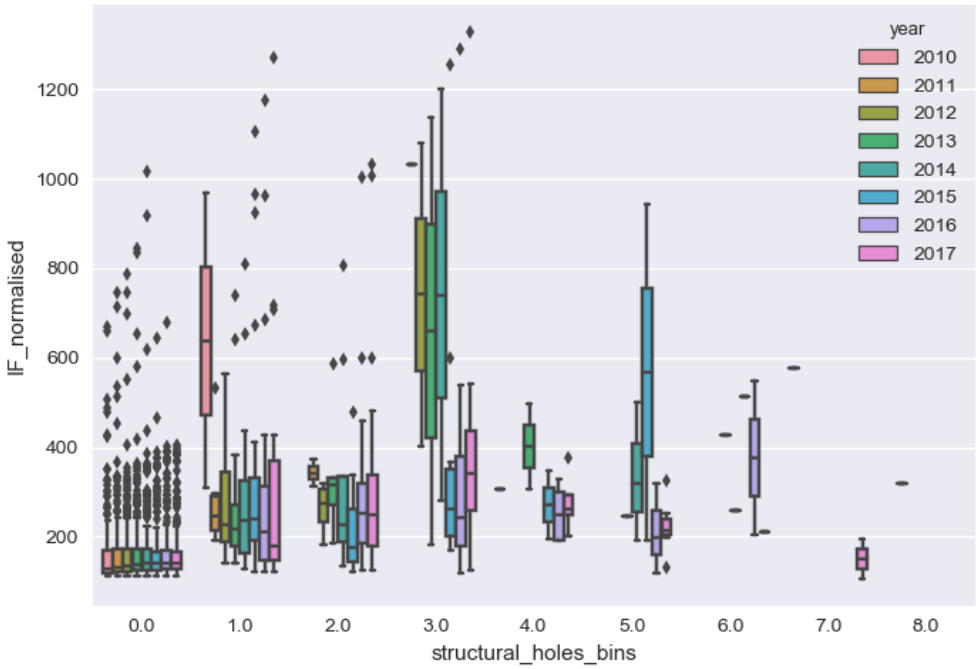

Figure 7.17. The box-plot for interdisciplinary authors as determined by department-based disciplines with $k \leq 30$ (as fewer points cause greater deviation) separated by time. The box-plot shows the interdisciplinary authors' structural holes vs. the interdisciplinary authors' impact factor normalised by the disciplinary trend from 2000-2010 to 2000-2017 (i.e. 8 time-periods). An overall negative trend can be seen between, and inconclusive trends can be seen within.

7.6.4. Model discussion

The structural holes measure provided a clear linear positive trend to corroborate Hypothesis 7.1. However, the reasoning as to why this would be better for interdisciplinary authors does not hold. Therefore, the model is only validated when having structural holes is an advantage. It is not validated to interdisciplinary authors having this structural advantage.

Hypothesis 7.2 is rejected as the findings were statistically insignificant. However, there was a positive trend, suggesting that it is possible that interdisciplinary researchers gain additional benefits from structural holes in comparison to disciplinary nodes.

7.7. Strength of weak ties

Granovetter's (1973) seminal work on 'the strength of weak ties' suggests that people with weak links will be heterophilious. This suggests that a situation where heterogeneous knowledge is desirable, weak links are the best at providing these.

The phenomenon suggests that the minimum cut in automatic community detection will be in these weak links. It also suggests that these weak links will have the highest betweenness scores (although these apply to nodes and not the links).

The strength of weak ties was originally only intended to highlight weak links as serving as bridges to different communities. However, with the concept of cross-fertilization, cross-functional teams, and horizontal organizational structures, there is ample evidence to suggest that weak links (as they were conceptualised) should serve as providing creative effectiveness and contributing to innovativeness. Many studies have focused on this and have confirmed this. Other studies have found the opposite.

The lack of consensus on the matter merits further investigation. This can be done by treating links as entities that can contain information, much like a node can. That is to say that a link can contain a total number of citations that all research between the two connecting nodes have collaborated on. Through this, a quantitative view of the strength of weak ties can be achieved.

As such, the strength of ties is going to be measured according to the definition of strength as per Freeman, White et al. (1992).

$$tie\ strength = number\ of\ collaborations\ between\ nodes\ i\ and\ j \quad (7.12)$$

$$s_{link} = A_{i,j} w_{i,j} \tag{7.13}$$

This allows ties themselves to be investigated in the same way degrees can be investigated. That is to say, it is possible to investigate whether repeat collaborations provide a net benefit to metrics of success. The following hypotheses should be tested.

It is important to note that the hypotheses described are node centric, which do not apply to ties. Therefore, the following bespoke hypotheses are proposed to be conducted on the University of Bath co-authorship network 2000-2017.

Hypothesis 7.3: The weight, w_{ij}, of interdisciplinary links, E_{inter}, is smaller than the weight of disciplinary links, E_{intra}.:

$$H_0: \mu\left(w_{ij_{inter}}\right) \geq \mu\left(w_{ij_{intra}}\right)$$
$$H_A: \mu\left(w_{ij_{inter}}\right) < \mu\left(w_{ij_{intra}}\right)$$

Hypothesis 7.4: The time-averaged weight of links, $\overline{w_{ij}}$, is negatively correlated to the time-averaged metrics of success, $\overline{Y_i}$.

$$\overline{Y_i} = \beta_0 + \beta_1 \overline{w_{ij}} + \alpha_i + \overline{u_i}$$

$$H_0: \beta_1 \geq 0$$
$$H_A: \beta_1 < 0$$

Where β_0 and β_1 are constants, α_i is the unobserved heterogeneity, and $\overline{u_i}$ is the time-averaged idiosyncratic error.

7.7.1. Model validity

Hypothesis 7.3 establishes whether interdisciplinary links are associated with weak ties. It can be established by performing a simple t-test, and does not require a panel approach. Therefore, the analysis is performed on the full 2000-2017 dataset.

The results (see Table 7.14 clearly show that there is a statistically significant difference in disciplinary and interdisciplinary links. The null hypothesis is rejected, and the alternative hypothesis is corroborated. Interdisciplinary links in this dataset have weaker weights.

Table 7.14. T-test statistical analysis of link weights.

The department-based disciplinary links mean value, μ $\left(w_{ij_{intra_{department}}} \right)$	2.6537
The department-based disciplinary links mean value, μ $\left(\mathbf{w_{ij_{inter_{department}}}} \right)$	1.7239
T-value:	9.4629
P-value:	4.0778e-21
The content-based disciplinary links mean value, μ $\left(w_{ij_{intra_{content}}} \right)$	2.7576
The content-based disciplinary links mean value, μ $\left(\mathbf{w_{ij_{inter_{content}}}} \right)$	2.0839
T-value:	7.7869
P-value:	7.9265e-15

Hypothesis 7.4 establishes whether the quality of output is affected by the weight of links. As the measure is centred on link, it cannot easily be associated with funding or the degree of a node. The only applicable metric of success that can be used is the impact factor. However, it is averaged as the weight of links is equivalent to the number of papers.

A weak positive regression was found as shown in Table 7.15. This corroborated the null hypothesis and rejects the alternative hypothesis. This suggests that the concept should be the strength of strong ties.

Table 7.15. The strength of weak ties linear regression results. The coefficient, β_1, is 0.1370 corroborating the null hypothesis and rejecting the alternative hypothesis. This is a statistically significant result.

OLS Regression Results

Dep. Variable:	mean_IF	R-squared:	0.162
Model:	OLS	Adj. R-squared:	0.162
Method:	Least Squares	F-statistic:	1272.
Date:	Thu, 26 Apr 2018	Prob (F-statistic):	9.15e-255
Time:	15:37:35	Log-Likelihood:	-11210.
No. Observations:	6578	AIC:	2.242e+04
Df Residuals:	6577	BIC:	2.243e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
weight	0.1370	0.004	35.659	0.000	0.129	0.145

Omnibus:	6549.227	Durbin-Watson:	1.650
Prob(Omnibus):	0.000	Jarque-Bera (JB):	767598.327
Skew:	4.591	Prob(JB):	0.00
Kurtosis:	55.118	Cond. No.	1.00

7.7.2. Discussion

The strength of weak ties has been a contentious measure, as it is based on sound principles, but assumes that distinct bridges exist between communities, although it is more likely that there are many bridges between different communities (Barabási and Pósfai 2016).

This analysis has shown that average weight of links between different disciplines are less than they are within disciplines. However, this analysis also finds there is a statistically significant trend to increase the quality of publications with repeat publications.

Therefore, for this dataset, the concept of ‘the strength of weak ties’ is rejected.

7.8. Chapter discussion

This chapter set out to adapt five models to include disciplinary and interdisciplinary authors. These models were then tested to see if they held and to establish whether there was a difference between disciplinary and interdisciplinary nodes. Observing such a difference would have served as the foundation to understanding why the differences exist and in turn help develop a model that can identify individuals who enable and sustain IDR.

Two hypotheses were designed to establish this. Hypothesis 7.1 was designed to establish whether the model held. Hypothesis 7.2 was designed to establish how interdisciplinary authors differed from disciplinary authors.

Hypothesis 7.1 found that all the models had trends with the concepts. The Eigenvector/PageRank model was rejected, although as literature had mixed findings, this was not altogether unexpected. The strength of weak ties models was rejected, and was therefore not considered further as it was not node centric.

This helped establish that many of the models were in fact applicable to co-authorship networks and the wider collaboration networks that they represent. That these were also applicable when hard boundaries were drawn around a research organisation represents a minor contribution to knowledge.

Having established that these models held (with the exception of the strength of weak ties), it was possible to determine whether any differences between disciplinary and interdisciplinary nodes existed. Hypothesis 7.2 was designed to establish this.

With the exception of the PageRank content-based disciplines, Hypothesis 7.2 was rejected across all the models. This means that with the exception of the PageRank content-based disciplines, no differences between disciplinary and interdisciplinary authors were found.

Delving further into the PageRank content-based disciplines provided many questions. The disciplinary and interdisciplinary had positive trends with the impact factor when the impact factor was not normalised. This may be an indication that this was driven by hubs of excellence. However, do these hubs consist of a mix of disciplinary and interdisciplinary people, and if so in which discipline definition? The fact that interdisciplinary authors in content-based disciplines provided greater benefit to interdisciplinary people is an indication that organic interdisciplinarity hubs provide the greatest benefits. These could be an indication of highly heterogeneous knowledge, but with low barriers (e.g. Mechanical Engineering staff from different parts of the department working on a single problem – this provides all the benefits of IDR, but with fewer of the drawbacks).

However, the question as to why this is not seen in either the degree or betweenness models is not explainable with the current research.

A study to explain why this could be occurring needs to be conducted.

Furthermore, that a difference was only discovered in one of the models very specifically does question the credibility.

Contribution to knowledge:

In testing the overall chapter hypothesis: "*SNA models show that there are differences between disciplinary and interdisciplinary authors.*" By virtue of Hypothesis 7.2 being rejected in almost all cases, the chapter Hypothesis is rejected. This has important implications as it suggests that the premise that there are disciplinary and interdisciplinary archetypes is refuted. The represents an original contribution to knowledge, that outlines part of the reason why traditional networks analysis is unsuitable to investigate IDR.

To overcome this weakness, it is necessary to shed the disciplinary and interdisciplinary authors paradigm, and to consider different types of links. This would reduce the overall number of assumptions in the analysis and accepts that there are no such thing as interdisciplinary authors, but just interdisciplinary links. It would be a more realistic representation of the reality of the situation (i.e. a person is classified to a particular discipline, and can be connected to any other individuals in any other discipline).

This would require a framework that is able to consider M^2 types of links (M is the number of disciplines). If this were applied individually to create M^2 networks, then every network would be very sparse, and the networks analysis would be meaningless. It is therefore necessary to create a framework that is adequately able to relate each of these links to each other, even though they are not equivalent.

Such a framework falls under the field of multilayer networks analysis. This was deemed to be the only reliable approach that can minimise the number of assumptions and provide a more effective way of analysing IDR through the use SNA.

This recommendation can be considered a contribution to knowledge, as it advances the knowledge of SNA seeking to investigate IDR.

Furthermore, as the models were correlated to the impact factor, but this provides no distinction between disciplinary and interdisciplinary outputs. This compounded with the difficulty in measuring IDR, makes this study and measure unsuitable. The future degree is therefore a better operational definition for enabling and sustaining IDR as it provides a direct measure of whether an individual collaborates more (enabling) and if this can show to hold predictively (sustaining) the research aim can be achieved.

7.9. Chapter summary

This chapter assumed that disciplinary and interdisciplinary authors existed as archetypes of authors. This work tried to establish if there was any difference between these two using SNA. Five different models were adapted to test this, and although the models themselves were corroborated (with the exception of the strength of weak ties), only a single case where differences occurred could be found.

Therefore, the archetypes concept is rejected (the case where there was a statistically significant trend is deemed insufficient). Furthermore, without distinguishing between interdisciplinary impact factors and disciplinary impact factors, this work correlates equal output. It is therefore necessary to develop a framework that can address these issues. A multilayer perspective could address these issues whilst simultaneously improving analytical resolution by considering the context of the IDR collaborations (e.g. Physics-Chemistry pair as the context between two collaborators from each discipline).

Chapter 8: Multilayer Networks Review and framework definition

As was demonstrated in Chapter 7, there is a need to extend the Networks framework to consider the different types of collaborations that can occur. This needs to go beyond whether a collaboration is disciplinary or interdisciplinary, and includes the context of the collaboration.

As such, it is necessary to adopt a multilayer framework that can adequately create such an analysis. Multilayer networks is a nascent field that is still developing its analyses, and this section establishes the various approaches found in literature.

It first establishes the different types of multilayer networks that can be adopted. It then outlines the intricacies of constructing such networks, and what benefits and challenges it provides. It then outlines the various centrality measures that may be of use. However, once these different centrality measures are understood, it becomes clear that none provide much additional benefit over traditional Networks analyses.

It is then established that multilayer networks has great potential in understanding the dynamics and evolution of a network. This is then chosen as the approach for identifying individuals who can enable and sustain IDR.

8.1. Approach

Multilayer networks have been the source of a lot of very recent research and the field is still in its infancy (Kivelä, Arenas et al. 2014). However, the research is very fragmented, applied in many different areas of research, and simply searching "multilayer networks" in Google Scholar yields over 35,000 results since 2014. This makes it difficult to review systematically. The same approach was taken as in Chapter 3, but a recent review by Kivelä, Arenas et al. (2014) provided a suitable foundation for the establishing the state of the multilayer networks. Relevant citations in the review were read and included in this review, and areas of interest were added upon.

8.2. Types of multilayer networks

Multilayer networks have been approached from a multitude of different disciplines. As such, many different names have been given to multilayer networks, each addressing some specific need, and that many of these terms are used interchangeably (Kivelä, Arenas et al. 2014).

Multilayer networks differ from traditional networks in that layers are integrated into the networks framework. Where a normal network is composed of two elements $G = (V, E)$, where V is the set

of nodes, and E is the links between all the nodes in V. Multilayer networks must include a term that accounts for what layer the node and links are, if we consider M layers, and define α as the source layer, and β the target layer, a multilayer network must be defined as $\Gamma = (V_\alpha, V_\beta, E_{\alpha\beta})$.

It is worth noting that other works have also integrated additional dimensions such as time, which requires the definition of an additional factor, which has been defined as a 'fundamental layer' (Kivelä, Arenas et al. 2014).

There are a number of different ways that implementing these factors would affect the mathematical framework of the multilayer networks, and this has given rise to many different types and names of multilayer networks. This research considers broad types: Multiplex Networks, and Network-of-Networks. As the field is still in its infancy, the definitions adopted here will likely differ from other works, but it provides a useful separation.

Multiplex networks

Multiplex Networks is defined as a series of network layers, each containing all nodes, but with different links. The term 'multiplex networks' originates from sociology (Wasserman and Faust 1994), and is one of the most thoroughly studied multilayer networks (Kivelä, Arenas et al. 2014). Sociology have recognised for a long time that approximating relationships to being equal in sociograms is a crude approximation, and better representations are needed (Krackhardt 1987, Padgett and Ansell 1993, Wasserman and Faust 1994). Transport networks have also employed the use of multiplex networks to recognise that transport routes may use different companies too, each of which is their own network (e.g. airline destination network layers constructing an air travel destination multiplex network (Cardillo, Gómez-Gardenes et al. 2013, Cardillo, Zanin et al. 2013), or public transport links in a city (Cozzo, Kivelä et al. 2013, Rombach, Porter et al. 2017).

The multiplex networks framework was proposed to accept different types of interaction individuals can have between them (e.g. multiplex communication networks could consist of different edge types for phone calls, emails, and in-person contact). This implies that there is a layer for every type of edge. These multiplex networks therefore consist of *M* layers with *N* nodes. All *N* nodes exist in all *M* layers. This suggests that multiplex networks are multilayer networks that contain different edges in each layer between the same nodes, and often the nodes are connected to their counterparts in different layers (Mucha and Porter 2010) as shown in Figure 8.1.

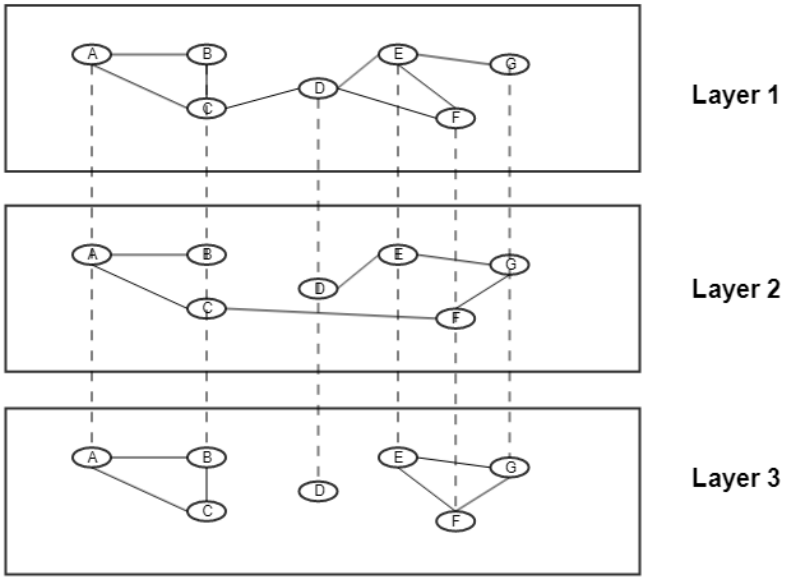

Figure 8.1. Multiplex networks consist of N nodes, which exist in each of the M layers. The edges in each layer are unique, and the layers are connected by virtue of the same nodes existing.

For these reasons, multiplex networks have also been called multi-relational networks, or edge-coloured networks (Coscia, Rossetti et al. 2013, De Domenico 2014).

Therefore, referring to the general form of multilayer networks, $\Gamma = G(V_\alpha, V_\beta, E_{\alpha\beta})$ can simply be rewritten as $\Gamma = G(V, E_{\alpha\beta})_{M\times M}$. It is important to understand that layers are not synonymous with disciplines, and interdisciplinary links between two layers will exist within their own layer, as such there are $M \times M$ layers of $G(V, E_{\alpha\beta})$. Furthermore, elements within $E_{\alpha\beta}$ will be associated with nodes i and j. However, the multiplex framework cannot have different nodes connected across layers and requires $E_{\alpha\beta_{ij}} = 0$ if $\alpha \neq \beta$ and $i \neq j$.

Having defined the basic form of multiplex networks, it is possible to discuss how it is that these networks can be manipulated and analysed.

There are two main representations of multiplex networks in the searched literature: Tensor and supra-adjacency.

The tensor representation takes advantage of the general form and node alignment of multiplex networks to realise that it is a rank-4 tensor of size *N×N×M×M (De Domenico, Solé-Ribalta et al. 2013)*. However, it is worth noting that some studies have attempted to use tensor notation in other multilayer network types by adapting their network to fit the tensor notation (De Domenico, Solé-Ribalta et al. 2013). However, this is just adapting another multilayer network to being quasi-multiplex.

Tensor methods opens different avenues to analyse multiplex networks by taking advantages of tensor decomposition methods (e.g. Singular Value Decomposition, SVD) to successfully create

centrality measures (Kleinberg 1999, Kolda, Bader et al. 2005, Kolda and Bader 2006), or modularity analyses for community detection in multiplex networks (Kolda, Bader et al. 2005, Dunlavy, Kolda et al. 2011, Bonacina, D'Errico et al. 2015). Tensor analysis can be directly applied to multiplex networks, but a new lens for analysis would have to be developed as traditional structural measures cannot easily be applied.

For instance, several papers have proposed that developing an Eigenvector centrality is possible using higher-order tensors (Solá, Romance et al. 2013). This is based on recent work that proves that the Perron-Frobenius theorem holds for higher-order tensors (Qi 2005, Chang, Pearson et al. 2008). However, the theorem assumes super-symmetry (i.e. $N \times \ldots \times N$ tensors), and therefore does not hold for multiplex rank-4 tensors of size $N \times N \times M \times M$.

It is difficult to apply traditional network measures based on square-matrices to tensors without super-symmetry, and bespoke measures need to be created.

Therefore, most established papers have proposed centrality measures based on flattening multiplex networks, which ultimately becomes a supra-adjacency approach (Solá, Romance et al. 2013).

Supra-adjacency methods (also known as super-adjacency) involves creating a new adjacency matrix that contains the information within and across layers (Cozzo, Kivelä et al. 2013, Gomez, Diaz-Guilera et al. 2013, Sole-Ribalta, De Domenico et al. 2013, Sánchez-García, Cozzo et al. 2014). The result is $NM \times NM$ matrix. However, this means that multilayer networks aggregate the various edges into the same, which results in the loss of information.

The use of supra-adjacency matrices is by far the most wide-spread approach for multilayer matrices, ranging from multiplex studies (Cozzo, Kivelä et al. 2013, Gomez, Diaz-Guilera et al. 2013, Sahneh, Scoglio et al. 2013, Sole-Ribalta, De Domenico et al. 2013, Wang, Li et al. 2013, Radicchi 2014) to MDM (Maurer 2007). Notable works have focused on understanding the effect that layers have on the spread and diffusion of properties (Gomez, Diaz-Guilera et al. 2013, Sahneh, Scoglio et al. 2013, Sole-Ribalta, De Domenico et al. 2013, Wang, Li et al. 2013). These use the supra-Laplacians to show that spread is faster through layers than they are in flat networks. The supra-Laplacian networks are obtained by getting the Laplacian of the supra-adjacency matrix.

Aggregated multiplex networks into traditional networks is another commonly taken approach. This differs from supra-adjacency in that it remains a $N \times N$ matrix. As multiplex networks contain different links, it is necessary to provide a mechanism as to how the different edges can be combined. A linear superposition with specific layer weights (e.g. differing weights between disciplinary and interdisciplinary layers) has been commonly used (De Domenico, Solé-Ribalta et al. 2013, Gomez, Diaz-Guilera et al. 2013, Battiston, Nicosia et al. 2014).

This traditional network would therefore consider the multilayer aspect by altering weights. If the layers were not to be weighted in the University of Bath co-authorship network, there would be no difference between the aggregated network, and the network analysis provided in Chapter 7.

The difficulty with applying weights is that there are only three methods of defining them. The most robust method would be some metric that can be collected alongside the source-data that would explicitly define the weights (e.g. the disciplinary confusion matrix or layer degree as a measure of layer centrality). This type of data would undoubtedly be difficult to find however. Otherwise, the weights could be inferred based on certain network measures (e.g. network community structure overlap (Cai, Shao et al. 2005, Rocklin and Pinar 2013)). Finally, the layers could be defined on some other metric, for instance, the Zachary Karate Club network was split and weighted on eight bespoke relational aspects (Zachary 1977). This has its obvious problems (the relational aspects chosen are not a complete representation of relationships, and any weighting needs to be validated for rigour).

This approach of course does not need to include all layers, and can instead choose a specific set of layers to aggregate (Corominas-Murtra, Fuchs et al. 2014). This could be useful to determine the structure of disciplines of interest (e.g. STEM).

Network-of-Networks

Network-of-Networks is defined as a series of network layers, with a unique set of nodes associated with each layer, and links existing between different nodes of the same layer and links existing either between the layers themselves, or between nodes of different layers. This would include studies on interdependent networks, which have been used to study the cascading power-failure that blacked-out Italy in 2003, caused by a computer-power interdependent networks failure (Buldyrev, Parshani et al. 2010, Ellinas, Hall et al. 2014).

Network-of-Networks have also been utilised in many studies and date back to 1973 (Craven and Wellman 1973). However, it is more useful as a concept than it is a mathematical framework and has been discussed in the context of electrical networks (Pahwa, Youssef et al. 2014), smart grids (Bompard, Han et al. 2014), and transport (Morris and Barthelemy 2012, Morris and Barthelemy 2014).

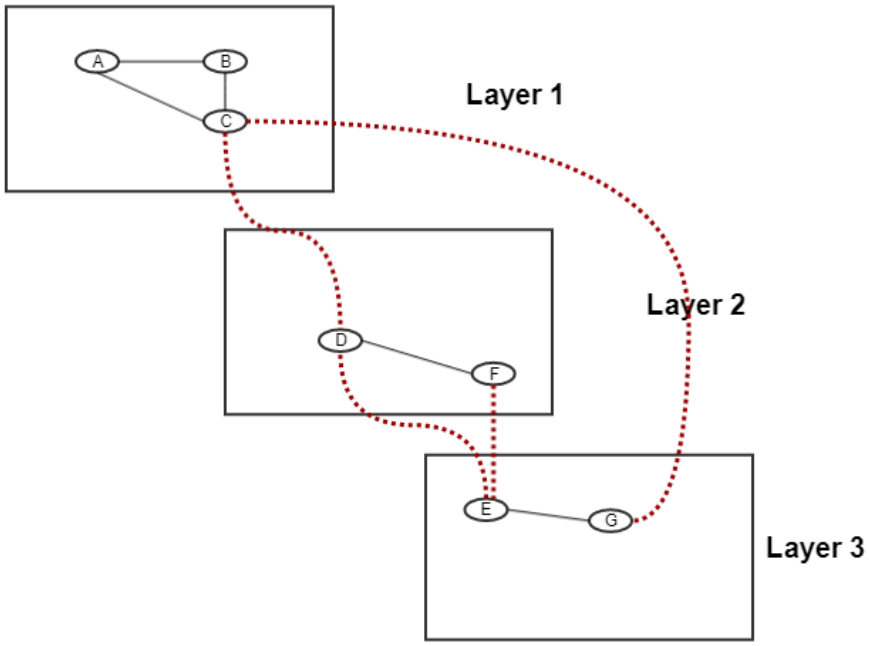

Figure 8.2.Network-of-Networks is composed of traditional networks within layers. However, the layers are connected to each other, either via node links, or links connecting layers themselves.

However, these types of networks are difficult to manipulate. Nodes in every layer, intra-layer links, and inter-layer links between two specific layers each require their own links. If creating a vector of the overall multilayer structural measure were desired, then network-of-networks approaches would be forced to introduce a bespoke weighting to the various link entities. This is difficult and would require validation for the weights used. If no weighting was used, then network-of-networks would be mathematically equivalent to traditional networks and no further benefit could be gained.

Therefore, as multiplex networks offer greater flexibility and rigour, they will be used over network-of-network approaches in this research.

8.3. Importance of multilayer networks

Multilayer networks have received a lot of interest, and it has been shown mathematically that multiplex networks exhibit different behaviour from equivalent aggregated traditional networks in several different studies (Gomez, Diaz-Guilera et al. 2013, Sole-Ribalta, De Domenico et al. 2013). By analysing the supra-Laplacian dynamics, it was found that multiplex diffusive processes occur in shorter time-frame than traditional networks (Sole-Ribalta, De Domenico et al. 2013).

In epidemiology, it has been shown that multiplex connectivity can alter the critical values for epidemic spread (Cozzo, Kivelä et al. 2013, Granell, Gómez et al. 2014). In a similar study, showed that either layer can dominate the overall dynamics of a layer, but that individual nodes can shift

the dynamics (Sahneh and Scoglio 2014). Granell, Gómez et al. (2014) use the Microscopic Markov Chain Approach (MMCA) to show that competing layers affect the spread of epidemics.

8.4. Datasets

The previous section described the different types of multilayer networks that have been studied. This section describes the data requirements. There are unsurprisingly few datasets available that can be reliably used to construct true multilayer networks. Many datasets focus on creating datasets that contain data within layers, but either fail to establish the connectivity between the layers, or the layers chosen are not a complete set (e.g. Florentine familial ties in the Medici era (Breiger and Pattison 1986, Cozzo, Kivelä et al. 2013)).

It has been described a weakness in most datasets that inter-layer strengths have not been defined (Kivelä, Arenas et al. 2014). This may not seem that applicable to multiplex networks for co-authorship networks. However, when one considers that information, knowledge, or inspiration may more readily be drawn from one field than another, its becomes obvious that layer links have not yet been considered.

8.5. Structural measures

Degree centrality in traditional networks is simply calculated and understood. This is true too of aggregated networks. However, there is some variation in approach. For instance, finding the degree of a node when only counting links that occur more than a certain amount of times across the different layers (Lytras, De Pablos et al. 2010, Bródka, Skibicki et al. 2011, Bródka, Kazienko et al. 2012). Other approaches have maintained analytical resolution by finding the aggregated degree of a given sub-set of layers, the comparison thereby providing valuable information (e.g. level of redundancy in a layer) (Berlingerio, Coscia et al. 2011, Berlingerio, Coscia et al. 2013, De Domenico 2014).

Other approaches have retained a layer separated representation and defined vector degree centralities as given in the following expressions for degree and strength respectively (Menichetti, Remondini et al. 2014).

$$k_i^\alpha = \sum_{i=1}^{N} \boldsymbol{A}_{ij}^\alpha \tag{8.1}$$

$$s_i^{\alpha} = \sum_{i=1}^{N} \boldsymbol{A}_{ij}^{\alpha} \cdot \boldsymbol{w}_{ij}^{\alpha} \tag{8.2}$$

Betweenness and closeness centralities (Szell and Thurner 2010, De Domenico, Solé-Ribalta et al. 2015) rely on shortest paths. However, the shortest paths in multiple layers is not a straight-forward concept. Aggregating the layers provides an obvious solution. However, explicitly taking into consideration the layers, it is necessary to understand whether the links between layers differ from intra-layer links (e.g. multiplex networks will have the same nodes, in co-author networks, this is akin to taking a step inside the mind of a person) (Cozzo, Kivelä et al. 2013, Kivelä, Arenas et al. 2014). In multiplex networks, if interlayer links (i.e. between the same nodes) are considered a step, then one can calculate the geodesic paths in the supra-adjacency network. The centralities for the layer specific nodes can then be calculated in the normal manner (Cozzo, Kivelä et al. 2013).

This assumes a classic random walker, where diffusive random walkers and maximal entropy random walkers are both suited to jumps, which easily support cross-layer jumps (De Domenico, Solé-Ribalta et al. 2014). In other cases, to achieve the multiplex node centrality, it can simply be summed into a single number (De Domenico, Solé-Ribalta et al. 2013, De Domenico, Solé-Ribalta et al. 2014). This method runs into the same difficulty in applying weights or cost of between-layer steps as determining the weight when aggregating layers.

The difference between intralayer and interlayer steps could be kept separate. For instance, finding the shortest path could be found by minimising the number of intra and inter layer lengths individually. This may result in a range of answers for any one pair, however, or the walks could be defined as a series of intra and inter walk path lengths. However, unless a specific weighting is applied to these, it is likely that no one shortest path can be attained, making the betweenness and closeness difficult to define (Sun, Han et al. 2011, Sahneh, Scoglio et al. 2013).

The interdependence of layers has been shown to be a very important feature in multilayer networks (Gomez, Diaz-Guilera et al. 2013). Finding the multilayer geodesic path can help define the **interdependence of layers** (e.g. the betweenness of layers) (Morris and Barthelemy 2012, Nicosia, Bianconi et al. 2013). Other methods of determining interdependence is to find the degree of global overlap between layers (Cellai, López et al. 2013). Other methods include comparing the various layer centrality sequences to determine interdependence (Nicosia and Latora 2015). Community commonality across layers has also been used as measure of interdependence (Melnik, Porter et al. 2014). There are a greater number of ways to determine interdependence in network-of-networks types (e.g. number of ties between layers, treating layers as nodes themselves).

Clustering coefficients are even more problematic to define. No clear best method has been established for weighted clustering; introducing a difficult to define weighting to a node's clustering calculation would make it even more difficult. A few attempts have been made (Magnani and Rossi 2013), but it is likely that an aggregated approach, or supra-adjacency approach would be the most robust.

The Eigenvector centrality needs to satisfy the Perron-Frobenius theorem to ensure that there is a unique vector associated with the largest negative Eigenvalue showing the relative importance of nodes. As was outlined, multilayer networks cannot satisfy this condition as no multilayer exhibits super-symmetry (Qi 2005). Solá, Romance et al. (2013) defined four different ways to calculate the eigenvector, by aggregating, keeping the eigenvector separate by layer, or by flattening the tensor into a supra-adjacency matrix.

However, the PageRank algorithm is more than suited to the task of determining relative importance centrality. The PageRank algorithm can deal with networks that are not fully connected by 'teleporting' (Lambiotte and Rosvall 2012). This was easily extended to layered networks, where the PageRank was calculated separately for every node in every layer.

Interlayer correlation is an important measure in multiplex networks. There two major approaches to interlayer correlation: degree-correlation, and edge overlap (Bianconi and Barabási 2001). Interlayer degree-correlation is important for the growth of multiplex networks as it is reasoned that if a correlation exists, then preferential attachment would incorporate preference from other layers (Nicosia, Bianconi et al. 2013, Nicosia, Bianconi et al. 2014, Nicosia and Latora 2015). This implies that the connectivity of nodes in layers are dependent on each other (positive node correlation), and would therefore create layer-pair hubs (Nicosia, Bianconi et al. 2013, Nicosia, Bianconi et al. 2014, Nicosia and Latora 2015).

8.6. Multilayer network evolution

The various approaches described have thus far described approaches to constructing, measures to analyse, and the dynamics describing the emergent nature of multilayer networks. To succeed in creating a model that can predict properties in multilayer networks it is necessary to understand them.

This requires in-part to understand the mechanisms responsible for the emergence of the structural properties seen.

The most basic multilayer network that could be created is to populate a set of layers with a set of nodes and to randomly assign connections between nodes as an Erdős–Rényi model. This provided

unrealistic network structures. The Barabási and Albert model has provided a simple mechanism by which realistic traditional networks can be grown (Barabási and Albert 1999). It was by identifying hidden mechanics in growth models that allowed Barabási and Albert (1999) to identify the importance of preferential attachment. This phenomenon has guided a lot of subsequent research (Barabási and Albert 1999, Barabási and Pósfai 2016).

Several growth models have been proposed. Non-linear preferential attachments were proposed as an improvement on the Barabási-Albert model (Krapivsky, Redner et al. 2000). A link selection model that selects random existing links and connects new nodes to either node that the link is connected to, thereby producing a scale-free distribution (Dorogovtsev and Mendes 2002). A copycat model has also been proposed that copies what other nodes are doing (Kleinberg, Kumar et al. 1999).

Bianconi-Barabási models attempt to improve on preferential attachment by suggesting that there is a node 'fitness' that determines how attractive a node is to receive new links (Bianconi and Barabási 2001). In cases where preferential attachment is driving the growth of the network, the degree of a node can be thought of as a fitness measure. However, it also opens the possibility of late-comers becoming major players.

However, each one of these models and variations include some measure of preferential attachment (whether it is direct or indirect) to create a realistic network structure. It is for this reason that the Barabási-Albert model (the first model that was able to recreate the scale-free property) has been cited over 30,000 times (verified on Google Scholar 24/06/2018). Optimization and game-theory approaches also suggest that preferential attachment is the rational choice (Fabrikant, Koutsoupias et al. 2002, Becker 2013).

However, each one of these models relies on the growth of the network, that is to say the addition of a node at every timestep. The sequential addition of nodes is a vital part of this dynamic.

Each of these models do not consider that two pre-existing nodes could create additional links between them. Link addition models have found that there is a preferential attachment that occurs on both ends of such links (Barabâsi, Jeong et al. 2002).

Equally, there are various different mechanisms that affect the overall process, such as aging (e.g. in collaboration networks, an author may have a prime) (Amaral, Scala et al. 2000), node deletion (e.g. an author switches jobs or retires) (Saavedra, Reed-Tsochas et al. 2008), and the number of links grows faster than the addition of nodes (e.g. number of scientific papers, and collaborators as shown in Chapter 5) amongst many different models.

The rich information that has been extracted from growth and evolution models has been invaluable in understanding networks. It is therefore important to develop similar models in multilayer networks. Not many studies have been performed.

For network-of-networks, Exponential Random Graph Models (ERGMs) have been extended to model the probability of local tie structures occurring. ERGMs have traditionally been thought as useful for social networks, as these are assumed to be locally emergent (Lusher, Koskinen et al. 2013). A network-of-network implementation of ERGMs was shown to successfully generalise French cancer research elites' collaboration network (Wang, Robins et al. 2013). Equally, ERGM was used to analyse special interest groups and identify social roles (Heaney 2014). Layer interdependency has been investigated based on degrees (Lee, Kim et al. 2012, Min, Do Yi et al. 2014), and has been extended to epidemics analysis (Funk and Jansen 2010). A similar approach has been taken in generating a multiplex network, but changing node labels to vary interlayer correlation (De Domenico, Solé-Ribalta et al. 2013).

Nicosia and Latora (2015) proposed two models based on simulated annealing to reproduce observed patterns in pairwise interlayer degree-correlations.

Nicosia, Bianconi et al. (2013) attempt to create a multiplex growth model based on linear preferential attachment. The preferential attachment is conducted layer-by-layer and the preferential attachment is based not only on the degree of the node in the layer in question, but also the degree in other layers. Two findings were given in the paper: the interdependency of the layers had a significant impact on the growth of the network and newer nodes are more affected by the interdependency. Nicosia, Bianconi et al. (2014) extends their model to include non-linear preferential attachment between the two layers. The paper succeeds in showing that altering the non-linear preferential attachment mechanisms across the layers changes the degree distribution, and layer-degree-correlations. This then provides a powerful framework to compare real networks to their growth results (e.g. if a negative degree-correlation is found in a real network, this could give an indication of how it is that the layers affect each other).

Kim and Goh (2013) provide similar approach where it is shown that the layer-pair's degree Pearson correlation coefficient changes significantly when coevolution parameter is altered. Generally speaking, the greater the extent of the coevolution, the greater the amount of degree-correlation, whilst also altering the degree distribution.

However, given the nature of multilayer networks and how many different formats there are, there are no growth models that focus on collaborations in particular. Furthermore, none of the studies provide conclusive evidence that a multiplex network structure has been successfully simulated. Finally, none of the studies have disseminated the various components and mechanisms through which multiplex networks can be formed.

These represent the greatest gaps in knowledge that need to be addressed in this research. These all provide the knowledge needed to understand how it is that individuals across disciplines collaborate with one another.

8.7. Framework definition

Multilayer networks provide the ability to differentiate between specific disciplines. This opens a lot of avenues to analyse a co-authorship network as it provides greater distinction within the model thereby increasing the analytical resolution.

Having reviewed the various approaches possible, it is possible to define the framework adopted in this research. This framework provides the foundation for analysing multiplex collaboration networks. It outlines what data is needed to create a multilayer network.

The framework here is the type of multilayer network that is best suited to analysing a multilayer co-authorship network. A network-of-networks approach is intuitive, but requires a set of assumptions on how different layers are connected to each other. Multiplex networks are more suitable as they provide a direct mathematical representation the multiplex structure by converting the network to its supra-adjacency form, and are more amenable to its tensor format.

After exploring both formats, the multiplex network format provided an easier and more rigorous implementation. The multiplex network, $\mathcal{G}$, is defined by its components.

$$\mathcal{G}(V, E, L) \tag{8.3}$$

Where V is the set of nodes that exist in every layer, E is a set containing the sets of edges for each layer α, $E^{\alpha \in L}$, and L is the set of layers. It useful to note that the convention adopted in this research uses superscript to define layer, and subscript to define node. However, even within the multiplex network framework, there are different ways to represent a collaboration network. There are two major approaches that could be taken to representing the layers. Either every discipline is its own layer as shown in the following expression.

$$L = \{l^0, l^1, \dots, l^M\} \tag{8.4}$$

Where L is the set of layers, l^α a specific layer, M the number of disciplines. This would imply that every node would have a presence in every discipline and interdisciplinary collaboration would be characterised by link overlap. That is to say that if two individuals are collaborating across different

disciplines, they would connected in both disciplines (layers). This is a multiplex network framework using a rank-3 tensor of the form $N \times N \times M$.

However, it is possible to define every layer as every type of link possible. This would result in $M \times M$ layers, as each discipline and discipline pair would be specific layer, as shown in the following expression.

$$L = \{l^{00}, l^{01}, \dots l^{0M}, l^{10}, \dots, l^{M0}, l^{M1}, \dots, l^{MM}\} \tag{8.5}$$

Where a layer is denoted by $l^{\alpha\beta}$, α and β are one of M disciplines. Where $\alpha = \beta$, the layer is disciplinary, and where $\alpha \neq \beta$, the layer is interdisciplinary. Unlike the other framework, individuals would have no activity in other disciplines as these would be represented exclusively in the interdisciplinary layers.

This latter way of defining layers provides a greater resolution as it distinguishes between different layer pairs. However, the former provided a highly simple and effective model that outperformed more sophisticated models. It also highlights the importance of individuals' presence in specific layers. For this reason, and as simplicity is the preferred approach, the former was adopted.

The adjacency is therefore given by the following expression.

$$A_{ij}^{\alpha} = \begin{cases} 1, & if\ E_{ij}^{\alpha} \in E^{\alpha} \\ 0, & if\ E_{ij}^{\alpha} \notin E^{\alpha} \end{cases} \tag{8.6}$$

The most important observation to make is that a node will exist in every layer (although they may not be active in that layer). This means that a node will almost be treated as a separate entity in every layer, which becomes an important feature in the analysis. This research names the overall presence of an individual a 'node', whereas the entities in individual layers are called 'node entities'. Therefore, a node is represented by all of its node entities.

Interdisciplinary links can be represented by the overlap. However, as most established networks measures are node-centric, it is useful to distinguish between disciplinary and interdisciplinary node entities. However, unlike in Chapter 7 where a threshold was needed for an entire node, here only node entities will be interdisciplinary and remains fairly assumption free. This simply requires a definition of which node entity represents the node's core-discipline.

The core-discipline is the layer element, $\alpha \in L$, that node *i* belongs to according to the classification method defined in Chapter 6. This is denoted as D_i (the discipline of *i*). Thus, a node in layer $D_i = \alpha$ will represent its disciplinary entity, whereas the node in $D_i \neq \alpha$ will represent one of its interdisciplinary entities (specific to layer α).

This culminates in the aggregate network being created and then split into layers as demonstrated in Figure 8.3

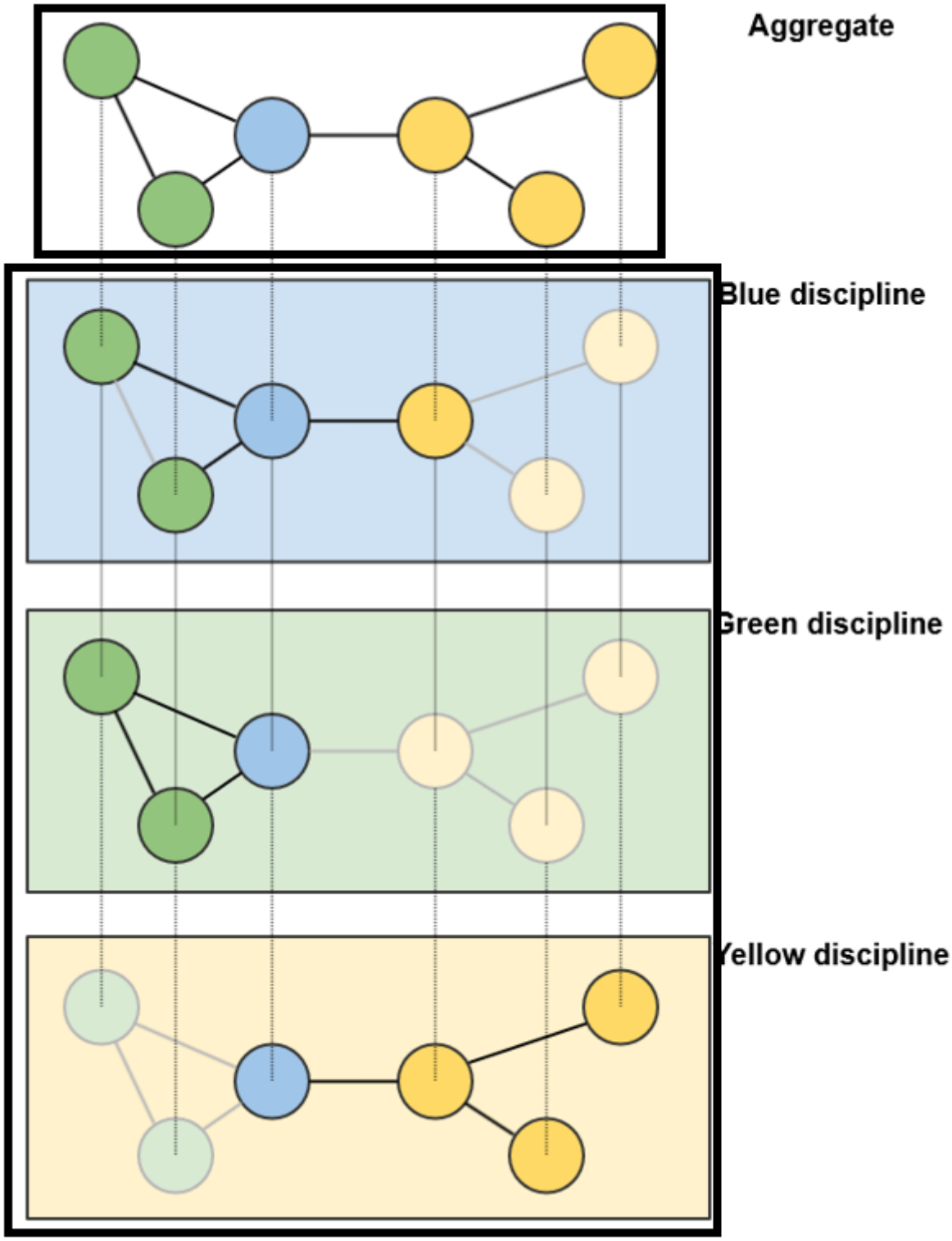

Figure 8.3. A conceptual representation of a traditional network (top box) and its counterpart multiplex network (bottom box). The networks are node aligned. The colours of the nodes represent the disciplines they belong to, each discipline having a layer. Any interdisciplinary links exhibit link overlap in both layers (e.g. the blue is connected to a yellow node in both the blue and yellow layers).

Finally, it is important to be aware that if a node has no links in a layer, this node is considered to be inactive in that layer. Node-layer activity is a vector of length *M* that is defined by the following expression.

$$b_i^\alpha = \begin{cases} 1, & if\ k_i^\alpha > 0 \\ 0, & otherwise \end{cases} \tag{8.7}$$

Multiplex node activity is the overall activity throughout the multiplex network given by the following expression.

$$B_i = \sum_{\alpha=1}^{M} b_i^\alpha \tag{8.8}$$

This provides the number layers a node is active in and can be thought of the node-layer degree.

Contribution to knowledge:

The section created a bespoke framework for multiplex collaboration networks seeking to investigate the effect of different classifications of individual such as individual's disciplines.

The framework is unique in that it identifies node classifications, but adopts a rank-3 tensor notation as opposed to a rank-4 tensor. This has a few pros and cons. The pros are that it circumvents 'the curse of dimensionality', wherein the more dimensions there are, the sparser the space, and more difficult it is to make any predictions. It also creates node entities, where every person has a network representation in every layer, which becomes vital in the predictive model this research produces. The cons are that it reduces the specificity of the analysis as specific interdisciplinary links are not identified. However, due to nature of multiplex networks, this information is not lost, but is rather shifted to the link overlap. In this respect, the rank-3 tensor provides all of the benefits with no loss of information (i.e. a rank-4 tensor representation can be constructed out of the information held in a rank-3 tensor representation).

This framework is an original contribution to knowledge that layers down the foundation for multiplex collaboration networks.

Chapter 9: Multilayer evolution in collaboration networks

Having defined how it is that multilayer networks are implemented in this research, it is possible to define the approach taken to achieve the research aim. The approach taken in Chapter 7 was to adapt and test previously reported successful models in a correlational study using the University of Bath co-authorship network. The study showed that traditional network approaches are unsuitable to differentiate between disciplinary research and IDR.

By expanding to the multilayer framework, several aspects are improved. There is greater differentiation to investigate interdisciplinary links by allowing different types of interdisciplinary collaborations. More importantly, it does not require a node to be disciplinary or interdisciplinary; a node's disciplinary and interdisciplinary collaborations are equally represented, thereby corresponding far closer to reality.

However, as multilayer networks are an emerging field, no models were found that are specific to collaboration. This makes it necessary to develop a new model. In order to do so, it is important to understand the factors that affect multiplex collaboration networks. Chapter 8 outlines that the growth of networks has provided invaluable insight into the nature of networks. It is the aim of this chapter to establish the foundations of such work so that such success may be repeated. By establishing the core mechanics by which multiplex collaboration networks form, it is possible to develop a model for how it is that individuals can enable and sustain IDR.

As multilayer networks is a nascent field, growth models can be developed upon and extended into a multilayer perspective. As described in Chapter 8, a few different approaches have been taken to develop a growth model for multiplex networks. However, no works centred on collaboration or co-authorship multiplex networks have been done and this represents a gap in knowledge that can achieve the research aim and provide significant insight into research policy (David 2013).

This chapter develops and validates a growth model, which in turn is validated as a predictive model that achieves the research aim. This model, and the findings leading up to the model, represent the major contributions to knowledge in this research.

As such, this chapter develops a model that can successfully recreate an exemplar multiplex collaboration network structure based on network properties alone. This in turn is found to have improved predictive capability over the models established in Chapter 7.

This chapter approaches the research as follows.

1. The approach defines how it is that the growth model can provide insight into real networks, and how it is that then achieves the research aim.
2. The methodology defines:

a. A series of multiplex network structural measures that adequately describe a multiplex network.
b. The modelling approaches.
c. The verification and validation methods.

3. It establishes the exemplar structure of the University of Bath multiplex co-authorship network based on the dataset 2000-2017. This is used to define the hypotheses to be tested on the growth models, and thereby serves as a historical data validation.
4. It proposes a series of growth models, presents the simulation results, tests the hypotheses, and discusses why the model behaves as it does.
5. The model is validated using predictive validation.
6. The model and its implications are discussed.

9.1. Approach

In order to be able to identify individuals who enable and sustain IDR, it is necessary to understand the mechanisms responsible for the emergence of the structural properties (Newman 2010, Barabási and Pósfai 2016). Barabási and Albert (1999) suggest that the reason that networks had not uncovered a method to simulate realistic networks prior to their study is because the Erdős-Renyi model did not take into consideration that networks grow and that links prefer nodes with more connections (Barabási and Albert 1999, Albert and Barabási 2002, Barabási and Pósfai 2016).

With multiplex networks, the model developed in this research uncovered further mechanisms that are needed to simulate a realistic multiplex network.

However, in order to perform this research in a rigorous manner, it is necessary to define a simulation verification and validation model. As per the deductive research philosophy adopted, the validation should also be expressed as hypotheses to be able to easily discern the knowledge created.

Ultimately, the growth model is validated using historical data and predictive validation. This validated model is then analysed to understand the implications and how it can then be used to achieve the research aim. Therefore, the chapter hypothesis is defined as follows.

Hypothesis 9: The multiplex collaboration network model identifies individuals who enable and sustain IDR.

9.2. Methodology

Having defined an overall approach, it is possible to define a methodology. The methodology defines the measures necessary to establish what a multiplex structure ought to look like. It then provides a verification and validation model. Finally, it develops the method for the growth models' creation.

9.2.1. Multiplex measures

This section defines the structural properties that suitably describe multiplex network structures. No one measure is suitable to define such a complex network. As such, several measures are necessary to describe the overall aggregate topology, the topology on individual layers, and the topology across layers (these are subsequently referred to as the "multiplex structure"). Some of the measures proposed here are well-established measures in literature (Newman 2010, Barabási and Pósfai 2016) that have been adapted to also investigate multiplex aspects. The other measures were proposed in Nicosia and Latora (2015), which focus on measuring the features across the layers.

Degree centrality is the staple structural measure that has been extensively studied (see Chapter 4). This can be extended to investigate multiplex structures' individual layers as shown in the following expression.

$$k_i^{\alpha} = \sum_{j=1}^{N} A_{ij}^{\alpha} \tag{9.1}$$

This will yield a vector of length *M* for every node, or a vector of length *N* for every layer. The aggregated degree would be the equivalent of a traditional network degree, given in the following expression.

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{9.2}$$

It is important to note that the format adopted in this research: aggregating the network is not equivalent to summing the vector of degrees, as interdisciplinary co-authorships exhibit link overlap across nodes. In a reversal of the adage, the whole is less than the sum of its parts.

$$k_i \neq \sum_{\alpha=1}^{M} k_i^{\alpha} \tag{9.3}$$

This measure, centred on nodes, can then create a distribution, which provides an indication of the multiplex network structure on all layers and the aggregated network.

As was outlined in Chapter 8, each node has a core layer, D_i. The node entities can therefore be split into two different sets: disciplinary node entities ($D_i = \alpha$) and interdisciplinary node entities ($D_i \neq \alpha$). Therefore, to gain a fuller perspective on the structure, the following measures should also be included.

$$k_{i\ intra} = \sum_{j=1}^{N} A_{ij}^{\alpha=D_i} \tag{9.4}$$

$$k_{i\ inter} = \sum_{\alpha \neq D_i}^{M} \sum_{j=1}^{N} A_{ij}^{\alpha} \tag{9.5}$$

Disciplinary-interdisciplinary degree comparisons compare a disciplinary node entity's degree to the sum of the corresponding interdisciplinary node entities' degrees.

The $y = x$ line provides a separation on whether an individual has more interdisciplinary or disciplinary collaborators. A non-linear trend would show whether higher degree individuals produce more or less IDR in comparison to other nodes, and therefore helps in profiling individuals who enable and sustain IDR.

Degree-correlation provides further indication on the structure of the layers by establishing whether nodes are connected to similar nodes. If there is a positive correlation, the layer is assortative. This means that high degree nodes tend to be connected to other high degree nodes (e.g. Hollywood actors (Barabási and Pósfai 2016)). These types of networks typically have shorter average pathlengths (Barabási and Pósfai 2016). If there is a negative correlation, the layer is disassortative, making the layer a more "hub-and-spoke" structure. The following expression provides a way to measure the degree-correlation of an individual.

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j \tag{9.6}$$

This measure is best represented as a distribution to determine the overall degree-correlation on every layer as shown in the following expression.

$$\overset{\alpha}{nn}(k_i^\alpha) = \frac{1}{k_i^\alpha}\sum_{j=1}^{N} A_{ij}^\alpha k_j^\alpha \tag{9.7}$$

Multiplex node activity is a node-centric measure that shows the number of layers a node is active in. Active is defined here as whether a node has a degree greater than one in a specific layer, as shown in the following expression.

$$b_i^\alpha = \begin{cases} 1, & if\ k_i^\alpha > 0 \\ 0, & otherwise \end{cases} \tag{9.8}$$

Multiplex node activity is the overall activity throughout the multiplex network given by the following expression.

$$B_i = \sum_{\alpha=1}^{M} b_i^\alpha \tag{9.9}$$

This provides the number of layers a node is active in and can be thought of the "vertical" degree.

Layer activity on the other hand is the number of active nodes within layer.

$$N^\alpha = \sum_{i=1}^{N} b_i^\alpha \tag{9.10}$$

Layer-pair closeness is needed as a measure of how close two specific layers are. Many shortest path algorithms have been developed for this purpose (Solé-Ribalta, De Domenico et al. 2014, Solé-Ribalta, Gómez et al. 2016), but a simpler implementation based on overlapping node activity suits the purposes here (Nicosia and Latora 2015). Simply by summing the number of nodes that are active in both layers provides a simple and effective measure. It is worth noting that the way that the multiplex network is defined, overlapping node activity guarantees overlapping links, which is another measure of how closely paired two layers are. The following expression gives such a measure.

$$Q_{\alpha\beta} = \sum_{i=1}^{N} b_i^\alpha \cdot b_i^\beta \tag{9.11}$$

This is also summed for each layer to find the **layer closeness centrality**.

$$Q_\alpha = \sum_{i=\beta,\beta\neq\alpha}^{M} Q_{\alpha\beta} \tag{9.12}$$

Each multiplex network consists of node entities in every layer. Comparing two such structures can therefore be quite difficult. The approach taken in this research is to compare defining features of the various distributions (e.g. the exponents of the degree distributions) for each layer and constructing a distribution of these (e.g. the distribution of the exponents). The shape and values of multiplex distributions would then characterise the multiplex structure.

However, as the population for the number of layers is small, some smoothing is required. A Kernel Density Estimation (KDE) approach allows for the distribution shape to be predicted and circumvent binning issues (although introduces tuning issues). Therefore, both the histogram and the KDE is included in distribution plots.

9.2.2. Verification and Validation model

There are many different modelling approaches and methods. Modelling is an iterative process that seeks to articulate the problem, form a dynamic hypothesis, formulate the simulation, test and compare the simulation results to reference results, and finally formulate and evaluate policy based-on the newly formed knowledge (Sterman 2000).

Another approach would be through the use of an Agent-Based Model's (ABM's) process. ABMs consists of defining actors with a set of rules on how they behave by themselves, in interaction with other agents, and with the environment. One of the first ABMs modelled the segregation of neighbourhoods (Schelling 1969), resulting in the 2005 Nobel Prize in Economic Sciences. Since then they have become popular with the advent of widely available computational capabilities, and statistical physicists have started applying their methods to social problems (Chakrabarti, Chakraborti et al. 2007).

With computer simulations becoming more commonplace, various complex social systems have been studied, such as modelling the stock markets (Arthur, Holland et al. 1996), traffic jams (Eisenblätter, Santen et al. 1998), the size of wars (Cederman 2003), as well as many network-centred ABM models like Susceptible-Infected-Recovered simulations (Barzel and Barabasi 2013). ABM approaches would therefore be suitable for networks modelling.

In ABM literature, two types of knowledge can be gained through ABM simulations: integrative and differential understanding (Wilensky and Rand 2015). Differential understanding is trying to understand what behaviours in agents can lead to observed overall patterns. Such simulations have

tackled many different issues ranging from contagion failures (Parandehgheibi and Modiano 2013), to epidemics' spread (Barzel and Barabasi 2013). Differential understanding has been used in these simulations because they serve as a strong mechanism to understand emergent behaviour, and would therefore be appropriate for this study.

This research adopts the verification and validation model presented in Sargent (2013), shown conceptually in Figure 9.1. This closely resembles the process defined in Sterman (2000) and the differential understanding from ABM. The modelling process therefore consists of modelling a conceptual model, specifying a simulation model, implementing the simulation model, and experimenting on the model to get results.

'Model verification' defines the approach taken to verify that the simulation implementation is an accurate representation of the conceptual model. 'Model validation' is the most important aspect as it describes how it is ensured that the model is a valid representation.

The 'Real World' space in this research represents the University of Bath collaboration, whilst the theory of interest is how IDR occurs.

The 'System Theories' in this research represents the network models that describe how IDR occurs. By matching the simulation results to the 'Real World' data, a new model is proposed (in System Theories), whilst simultaneously achieving operational validation.

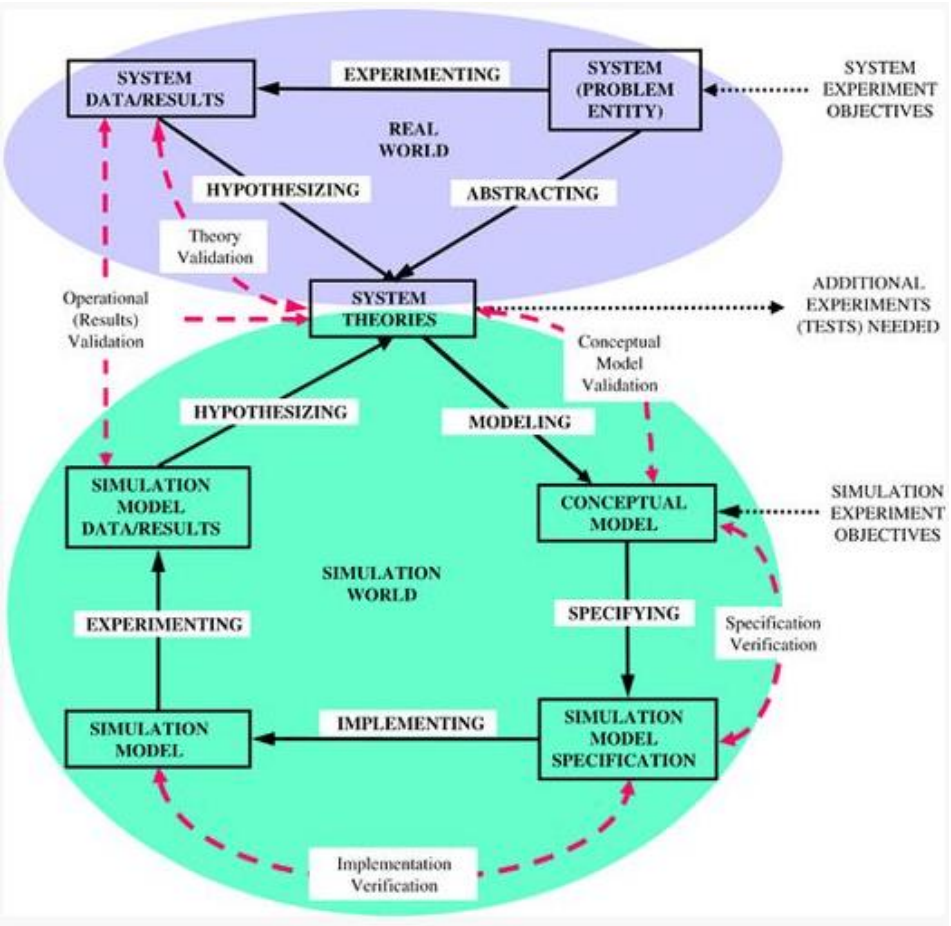

Figure 9.1. Verification and validation model from (Sargent 2013), originally from (Sargent 2001, Sargent 2013).

9.2.2.1. Modelling process

The modelling process consists of designing the conceptual model. The conceptual model in this research is a network growth model that has been extended to include multiplexity.

The models are built with the understanding that such a model can be either overfit or underfit. An overfit model in this case would be including highly complex rules and many different variables, which may make the model too specific for a particular organisation or time-frame. Therefore, the model intends to remain as simple as possible, not unlike the original Barabási-Albert model.

For a growth model, this means creating a minimalist model that accurately recreates the structure shown by the real multiplex network.

The conceptual model defines the creation of nodes, links, and layers. Nodes and links can both be added by any mechanism. The output of the models should always be a multiplex network, which can then be analysed using the multiplex measures defined (see 9.2.1).

9.2.2.2. Specifying process

Creating a specification is the capture of the model requirements. For research conducted by an individual, remaining agile was determined to take precedence, and the conceptual model remained the formal definition of the simulation.

9.2.2.3. Implementing process

Having decided upon a growth model to test, it is important to ensure that the implementation of these models corresponds to the conceptual model.

Verification of an implementation can be approached in several different ways. In larger projects, there tends to be peer-review of code. This is usually done by reviewing contributions to a project via version control platforms (e.g. GitHub, BitBucket). This ensures that code is readable, maintainable, and most importantly behaves correctly. Smaller projects require similar face validity (Xiang, Kennedy et al. 2005).

For models, such as network growth models, an equivalent approach is necessary to ensure that the implementation accurately represents the conceptual model proposed.

The following processes ensured that the implementation was verified in this research.

- Special attention was given to the time-stepping mechanism.
- The code and conceptual model were provided to Python-literate software developers (Sargent 2013).

- The results were run multiple times to ensure replicability and ensured that no consistent biases were found for a layer (Ormerod and Rosewell 2009).
- Tracing variables during modelling allowed bugs to be detected and corrected.
- Continuity testing was performed by changing probability inputs to ensure that representative changes occurred in the output.
- Degeneracy testing was also implemented to ensure that behaviours occur as expected in extreme values.

These verification methods along with attention to the implementation of the conceptual model ensured good results were obtained.

9.2.2.4. Experimenting process

The experimenting process consisted of testing the hypotheses using the multiplex metrics on the multiplex network created in the simulation. This ensured that the deductive philosophy is adhered to with regards to creating knowledge.

9.2.2.5. Validation process

Validation is arguably the most important step in modelling. The operational validation occurs in two different manners. Historical data validation requires the simulation to behave as the system does (Sargent 2013). Predictive validation consists of using the model to predict the system's behaviour using the historical longitudinal dataset. The validation should take the form of hypotheses tests. These are defined in the University of Bath multiplex co-authorship network section.

The conceptual model validation also took into consideration input validation. The input values should match to create a similar network to the real network.

The real network consists of 1,941 unique nodes with collaborators over 17 layers with more than 10 individuals on each layer. The rate of growth should match the real growth. Table 9.1 shows the increase in number of collaborators. Table 9.2 shows the increase in number of collaborators within their own discipline. Table 9.3 shows the increase in the number of interdisciplinary collaborators. It is important to note that these are not exclusively new collaborators to the University of Bath, but rather forming new collaborations, between new or old researchers.

Therefore, the models should ensure that the proportion of interdisciplinary growth is faster as the network matures, but slower initially.

Table 9.1. Number of new collaborators per year (all collaborators).

Year	New collaborators	Percentage increase
2011	266	25.732
2012	256	19.975
2013	295	17.365
2014	328	14.534
2015	341	13.602
2016	385	13.445
2017	676	14.564

Table 9.2. Number of new collaborators per year (disciplinary collaborators only).

Year	Additional disciplinary collaborators – all layers	Percentage increase
2011	167	25.891
2012	163	20.074
2013	169	17.333
2014	167	14.598
2015	178	13.577
2016	201	13.499
2017	222	13.136

Table 9.3. Number of new collaborators per year (interdisciplinary collaborators only).

Year	Additional interdisciplinary collaborators – all layers	Percentage increase
2011	99	29.290
2012	93	21.281
2013	126	23.774
2014	161	24.543
2015	163	19.951
2016	184	18.776
2017	454	39.003

To mimic these values, the growth models used the following parameters (where appropriate):

Table 9.4. Input parameters for the growth models.

N (final)	2295
$\mathbf{m_0}$	2
$\mathbf{m_1}$	2
$\mathbf{C_0}$	Tuned to achieve equal number of disciplinary and interdisciplinary links
M	17

9.3. The University of Bath multiplex co-authorship network 2000-2017

It is necessary to outline the results for real networks as this serves as the validation for the growth models. It also provides a lens through which we can discuss the growth model results. The multiplex network is created according to the defined framework (see Chapter 8).

The University of Bath multiplex network's layers represent its disciplines. The layers are all the disciplines that have been identified using the operational definitions as defined in Chapter 6. The nodes exist in all layers but are considered inactive if they have no co-authors. Every node has already been classified as being native to a core-discipline, D_i. If two authors have co-authored a journal paper, a link is added between their respective nodes on layers D_i and D_j. Where a node is active on a layer, its layer node entity exists. The multiplex network is node aligned.

As such, two different multiplex networks have been created and analysed. The department-based disciplines and the content-based disciplines form different multiplex structures. However, the trends agree over the two multiplex networks.

This achieves two things. Firstly, it defines an exemplar multiplex network that growth models can be validated to. The department-based network is chosen as the exemplar due to its more rigorous classification method. Secondly, it also provides rich information about the University of Bath collaboration. Furthermore, by comparing the two multiplex networks, further insight can be gained on content-based disciplines and department-based disciplines. It is also worth noting, that specific information and structural measures for each node could be picked out by focusing on any one layer or subset of nodes for specific policy decision. This is outside the scope of the thesis but represents a useful tool that has already been developed (see http://www.hultin.uk/visualise - to be released on 14/07/2018).

This section outlines the results for the degree distribution, disciplinary-interdisciplinary correlation, degree-correlation, node-layer activity, layer activity, and layer-pair closeness

providing a set of different measures that offer an overview of the multiplex structure. A series of hypothesis tests for the growth models are developed for each of these. These hypothesis tests determine the extent to which the growth models are valid to represent multiplex collaboration networks.

9.3.1. Department-based multiplex network

This section shows the results for multiplex structural measures for the department-based disciplines.

9.3.1.1. Degree distribution by layer

This section presents the results of the various degree distributions that were created from the multiplex network and defines hypotheses that seek to collectively answer Hypothesis 9.1.

> ***Hypothesis 9.1: The degree distribution of the model matches the degree distribution of the University of Bath multiplex co-authorship network.***

The aggregate degree distribution provides a relatively standard scale-free distribution with a very strong statistically significant correlation. The power-law coefficient is -1.87, which is a lower magnitude than many reported co-authorship trends. However, it is likely skewed by variation in the tail of the distribution.

This therefore suggests that the hypothesis for growth models should be defined as follows.

> ***Hypothesis 9.1(a) - The aggregate degree distribution produces a power-law relationship with an exponent between -1.5 to -2.5***

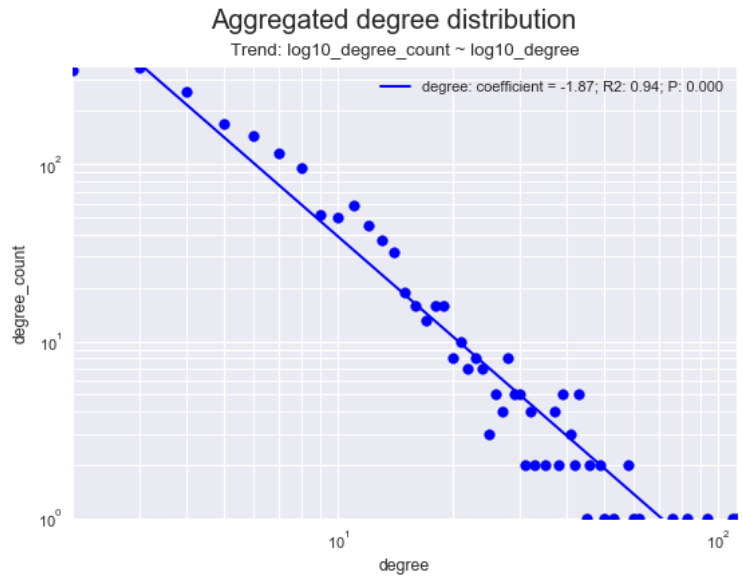

Figure 9.2. The aggregate degree distribution of the University of Bath department-based multiplex co-authorship network. This is equivalent to the traditional networks' degree distribution. It is statistically significant and has a strong correlation with an R^2-value of 0.94.

The layer degree distributions have been analysed on all layers using all node entities, disciplinary node entities only, and interdisciplinary node entities only, as seen in Figure 9.3, Figure 9.4, and Figure 9.5 respectively. A similar degree distribution to the aggregate network is seen throughout. It is worth noting that the interdisciplinary node entities' distributions contain a significant amount of variation and produce more statistically insignificant results. The 'politics, languages, and international studies' layer is not statistically significant in any of the plots; likely due to it only containing 17 node entities (resulting in 3 or 4 points in the distribution).

Hypothesis 9.1(b) - The degree distribution on every layer produces a power-law relationship using all node entities, disciplinary node entities only, and interdisciplinary node entities only.

Figure 9.3. The layer degree distributions of the University of Bath department-based multiplex co-authorship network. The majority of the distributions are roughly parallel. This includes all node entities.

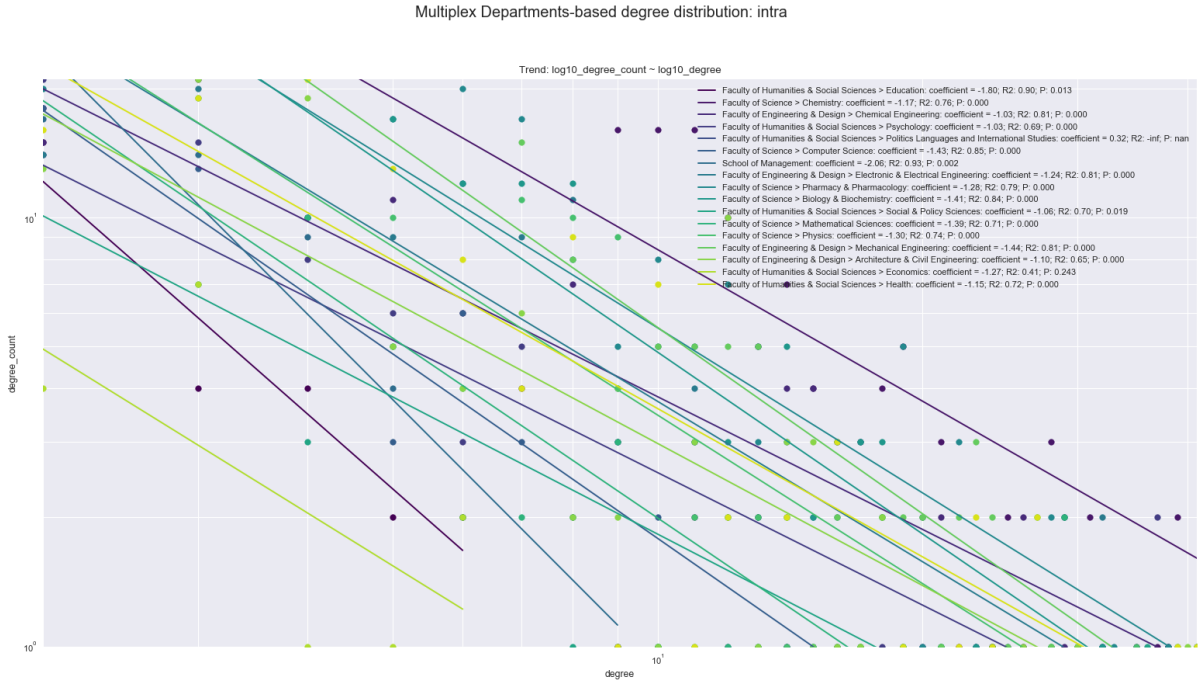

Figure 9.4. The layer degree distributions of the University of Bath department-based multiplex co-authorship network. The majority of the distributions are roughly parallel. This includes disciplinary node entities only.

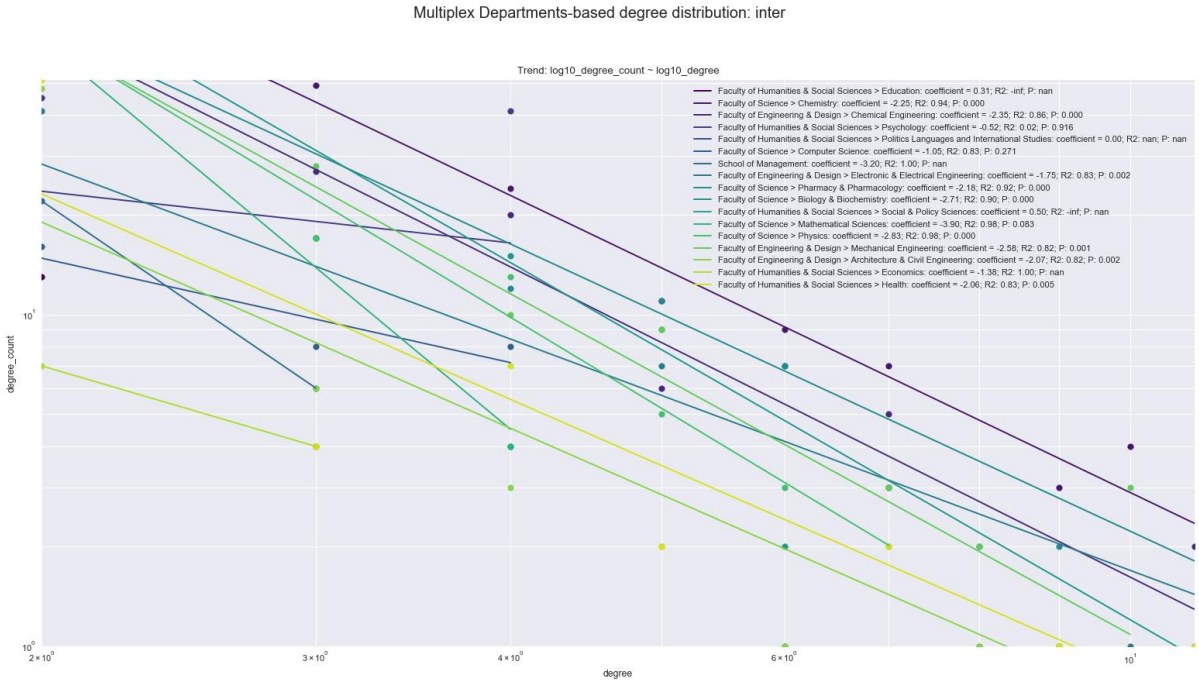

Figure 9.5. The layer degree distributions of the University of Bath department-based multiplex co-authorship network. The majority of the distributions are roughly parallel. This includes interdisciplinary node entities only.

The biggest difference between disciplinary and interdisciplinary node entities is that the number of collaborators is larger for disciplinary node entities logarithmically.

The statistically significant exponents form a distribution for all node entities, disciplinary node entities, and interdisciplinary node entities as shown in Figure 9.6. From this distribution, the following hypotheses can be formed.

Hypothesis 9.1(c) - The degree distribution on every layer, using all node entities, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than the aggregate exponent.

Hypothesis 9.1(d) - The degree distribution on every layer, using disciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than all the node entities' peak exponent.

Hypothesis 9.1(e) - The degree distribution on every layer, using interdisciplinary node entities only, produces power-law

exponents whose peak KDE density occurs at an exponent above the aggregate exponent.

Hypothesis 9.1(f) - The degree distributions' exponents are distributed as Gaussians that are skewed to the right as estimated by the KDE.

The most important aspect is the difference between disciplinary and interdisciplinary node entities' degrees. This suggests that there is a difference in the way that we conduct disciplinary and interdisciplinary research.

Contribution to knowledge:

The University of Bath Network 2000-2017 shows that there is difference between disciplinary and interdisciplinary node entities' structures. Given that node entities are representations of a single individual, the difference occurs within individuals. This means that the same individuals produce different network connections in IDR in comparison to disciplinary research, suggesting that it is not purely based on an individual, but also by the process.

This manifests itself in the network structure by having more interdisciplinary node entities with fewer collaborators that their disciplinary counterparts.

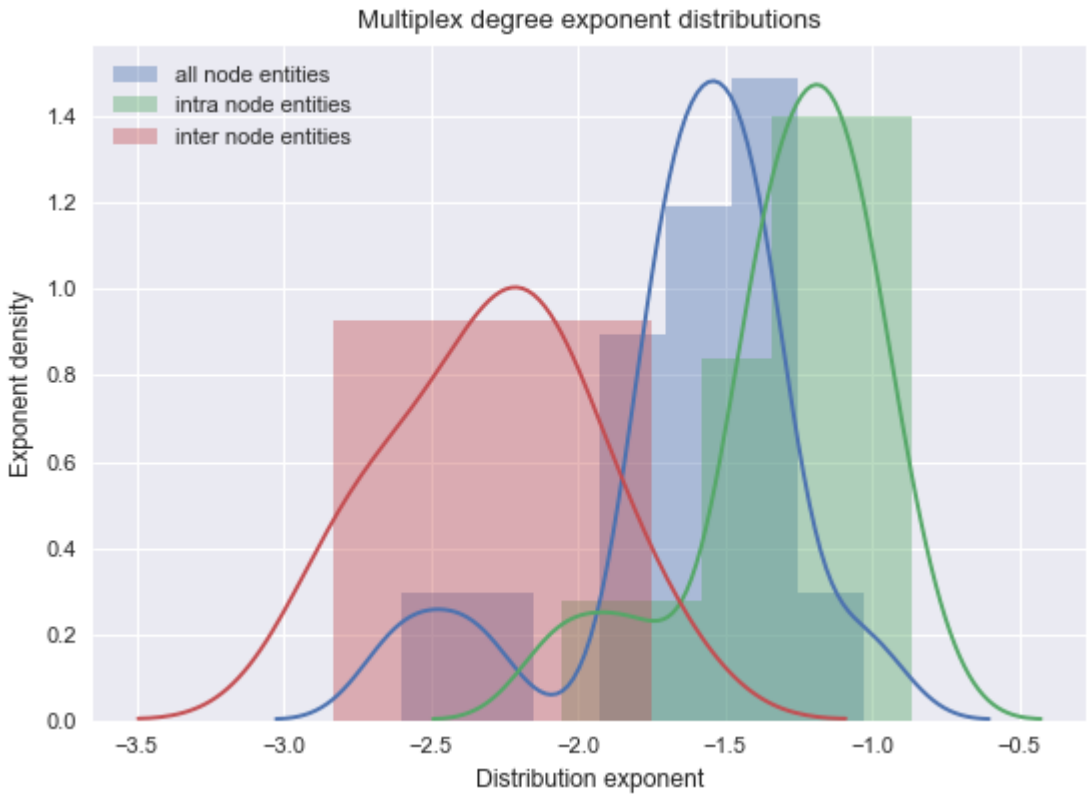

Figure 9.6. The layer degree distributions of the University of Bath department-based multiplex co-authorship network's all node, discipline only and interdisciplinary only exponents.

Ultimately, degree distributions have been a standard approach in almost all networks research because it is so useful. The implication of the power-law distribution is that the vast majority of people are poorly connected and very few people have a very large number of connections. This means that when trying to identify the individuals driving research, there are a few people who will be very central to a lot of research.

The fact that there is a difference in the structure of this distribution between disciplinary and interdisciplinary researcher when separating by disciplines provides an indication that there is a difference between the process of disciplinary research and IDR. The most apparent manifestation of this difference is that there are more poorly connected interdisciplinary researchers in a field than there are disciplinary. This finding is not wholly unexpected as it implies that IDR occurs often, but is unlikely to be sustained. Therefore, it becomes even more important to identify the individuals who seek to develop paradigms between two different disciplines. This is done by identifying the IDR node entities with the highest degree.

This could partially be done by identifying the individuals who have already overcome the barriers to IDR and are sustaining IDR (i.e. IDR entities who have a high degree in the 'IDR discipline').

This is a central tenet to the research, highlights the importance of node entities, and becomes an integral part of the growth and predictive models.

9.3.1.2. Disciplinary vs interdisciplinary degree regression

This metric can identify if there are any tendencies for individuals with a large number of connections to increase or decrease the amount of IDR conducted proportionally. However, due to the variation in individuals, a linear trend fits well (i.e. IDR is conducted proportionally on average).

Figure 9.7 shows that individuals clearly prefer collaborating within their own disciplines. This is an important finding, and a clear characteristic of real co-authorship and collaboration networks. Completely randomly connected networks would have a strong presence above the 1:1 line as there are more node entities outside their core-discipline than in it.

Therefore, the following hypothesis can be formed.

> ***Hypothesis 9.2: The disciplinary node entities' degrees are larger than the median of the sum of their counterpart interdisciplinary node entities' degrees.***

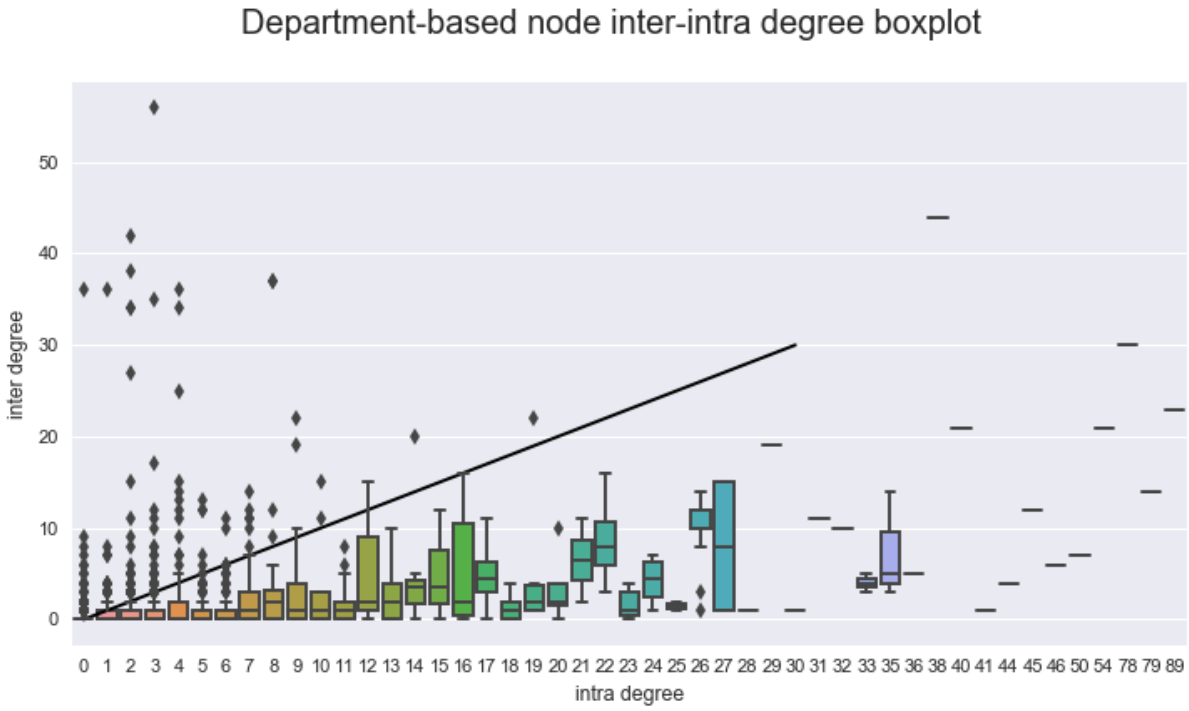

Figure 9.7. A box plot showing nodes' disciplinary (intra) degrees compared to the sum of their interdisciplinary (inter) degrees. The black line shows a 1:1 ratio of these. Note that the scale is not linear after a disciplinary degree of 33. There is a clear preference for individuals to collaborate within their own disciplines.

Whilst this measure is very simplistic. It provides an immediate perspective on the proportion of IDR to disciplinary research that occurs at the University of Bath 2000-2017. Despite the disincentive by REF to collaborate within disciplines since 2014, the majority of collaboration still occurs within disciplines (based on 2014-2017 co-authorships). Despite only one author being able

to take credit for a publication per department, most of the co-authorship occurring within Bath is still disciplinary. The disincentive may be manifesting in different ways; researchers could be publishing more solo papers, or may be seeking collaborations outside the University boundaries.

9.3.1.3. Degree-correlations

The degree-correlation shows the average degrees of a node's neighbours. This provides insight into how multiplex networks are structured, defining the next hypothesis.

> ***Hypothesis 9.3: The degree-correlation distribution of the model matches the degree-correlation distribution of the University of Bath multiplex co-authorship network.***

Figure 9.8 shows the degree-correlation for the aggregate network. Despite it not being statistically significant, it does show a slight positive trend that has been reported in scientific collaboration networks (Barabási and Pósfai 2016).

However, as it is not statistically significant, a hypothesis regarding the aggregate degree-correlation is not formed.

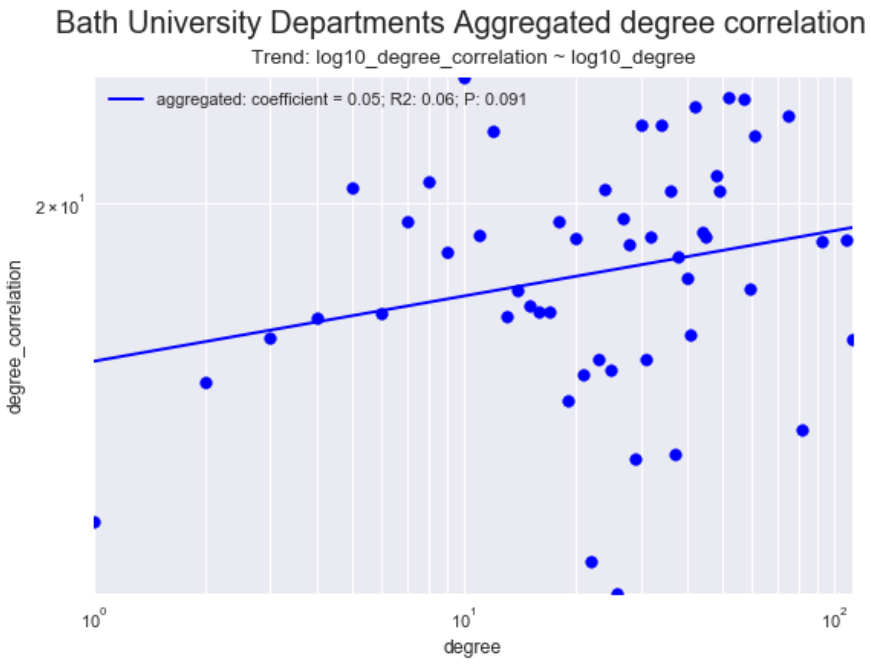

Figure 9.8. The degree-correlation for the aggregate network. Whilst it shows a positive trend, this is very poorly correlated and not statistically significant.

However, when the degree-correlation is done on individual layers, the trend reverses and becomes statistically significant, as can be seen in Figure 9.9. The degree-correlations therefore show that the multiplex structure has a different degree-correlation to traditional networks.

Contribution to knowledge:

A previously unseen phenomenon has been observed in the University of Bath 2000-2017 structure. Collaboration structures have been previously identified as having neutral degree correlations (i.e. there is no preference or dislike for highly connected individuals to collaborate with each other) (Barabási and Pósfai 2016).

When observing the overall network structure, this trend could not be rejected to the 0.05 level (although there is a small positive correlation that is significant to the 0.1 level). However, when broken into disciplines layers, a negative degree correlation (hub-and-spoke type structures) occur on all layers.

This means that a hub-and-spoke structure occurs within disciplines, but not overall. Therefore, different personal structures must be occurring on different disciplines (e.g. a researcher is a hub in one discipline and spoke in another, the aggregation would cancel these two out). This could be indicative that node entities play an important role.

Disciplinary node entities loosely match the distributions as can be seen Figure 9.10. However, interdisciplinary node entities' distributions are more statistically insignificant as shown in Figure 9.11. This is not merely a sample size issue as the degree distributions (see 9.3.1.1) are able to exhibit statistically significant results.

This behaviour has not been observed before, of which there are two aspects. The first is that the individual layers become statistically significant. The second is that there is an apparent trend reversal. The former can be defined as hypothesis.

Hypothesis 9.3(a) - Layers exhibit degree-correlation distributions with a power-law relationship with a negative exponent.

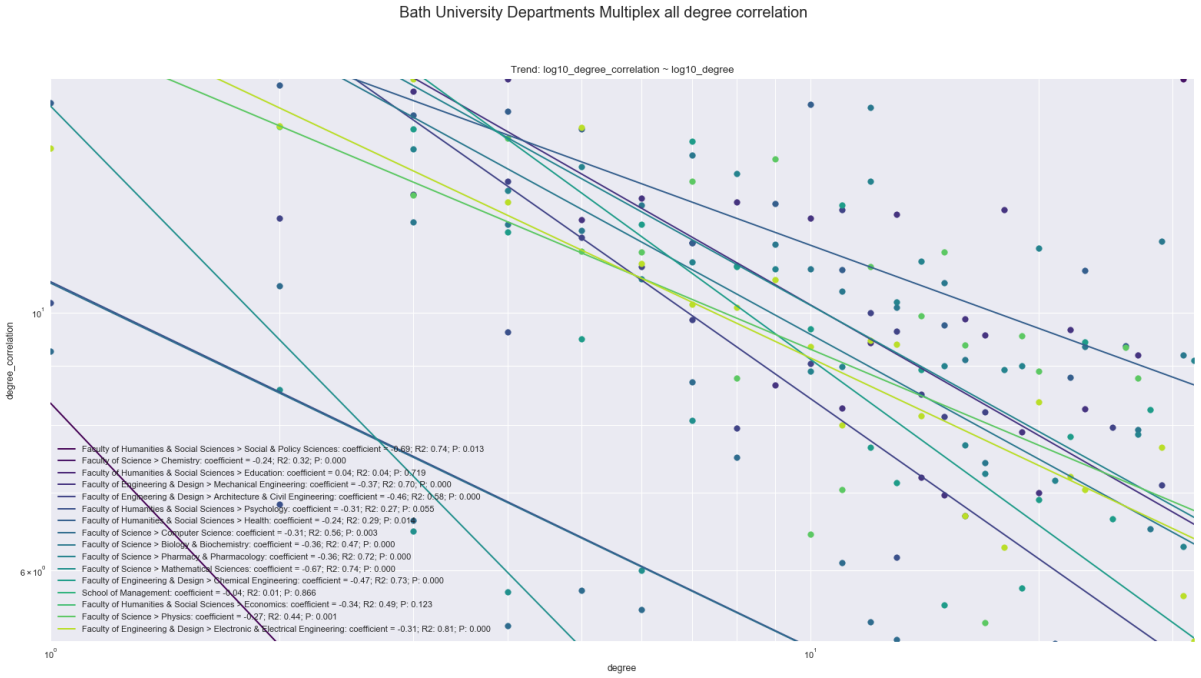

Figure 9.9. The degree-correlation distributions for the University of Bath department-based multiplex co-authorship network. This includes all node entities. Unlike previously reported studies, the collaborations within the layers are disassortative, and are mostly statistically significant (with the exception of Education, Psychology, Management, and Economics).

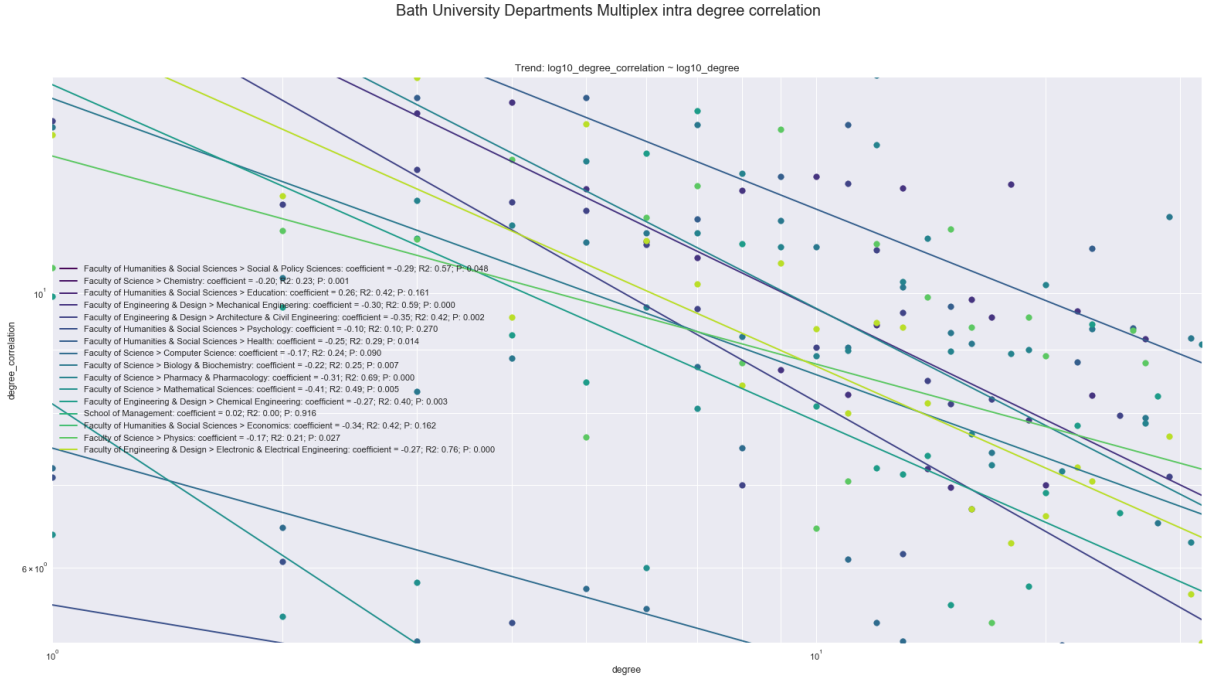

Figure 9.10. The degree-correlation for the University of Bath department-based multiplex co-authorship network. This includes disciplinary node entities only. The degree-correlations are disassortative.

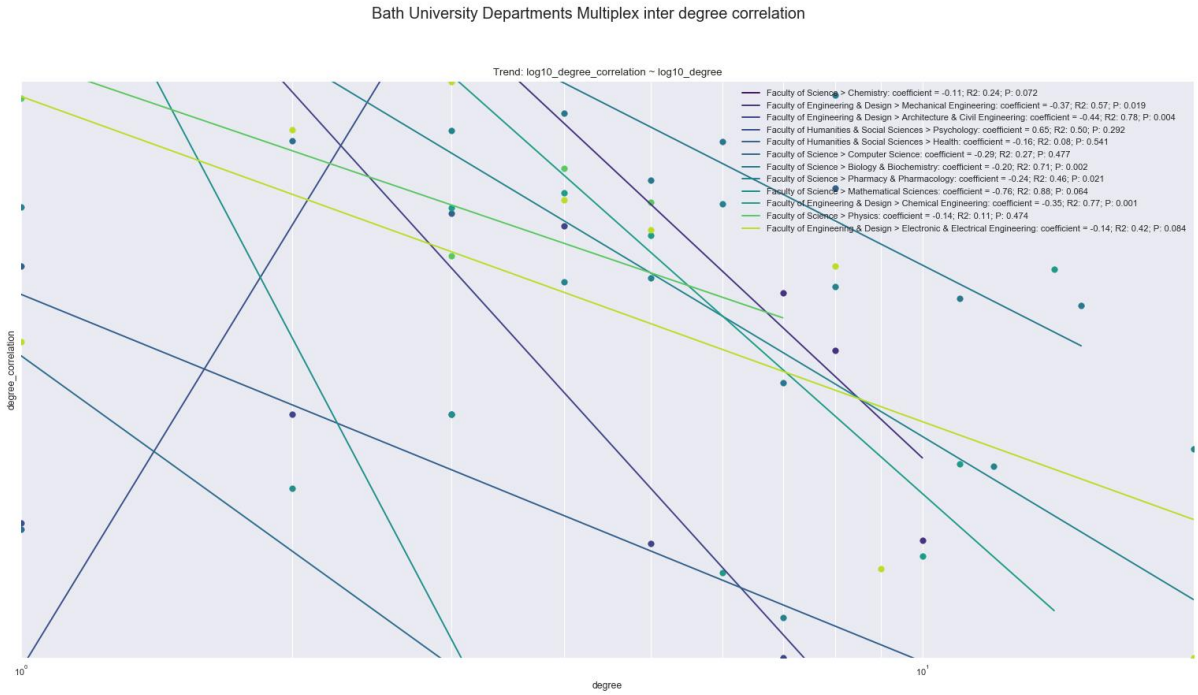

Figure 9.11. The degree-correlation for the University of Bath department-based multiplex co-authorship network. This includes interdisciplinary node entities only. The degree-correlations are disassortative. Many of the distributions are not statistically significant.

The statistically significant distributions all follow a power-law relationship, with the exponents in the layers being distributed as shown in Figure 9.12. The exponents form a Gaussian distribution skewed right. However, the exponents are very small in magnitude. The following hypothesis is therefore formed.

Hypothesis 9.3(b) - Degree-correlation distribution exponents exhibit a Gaussian distribution as estimated by the KDE skewed right with the peak density occurring at a value of $\gamma > -0.3$.

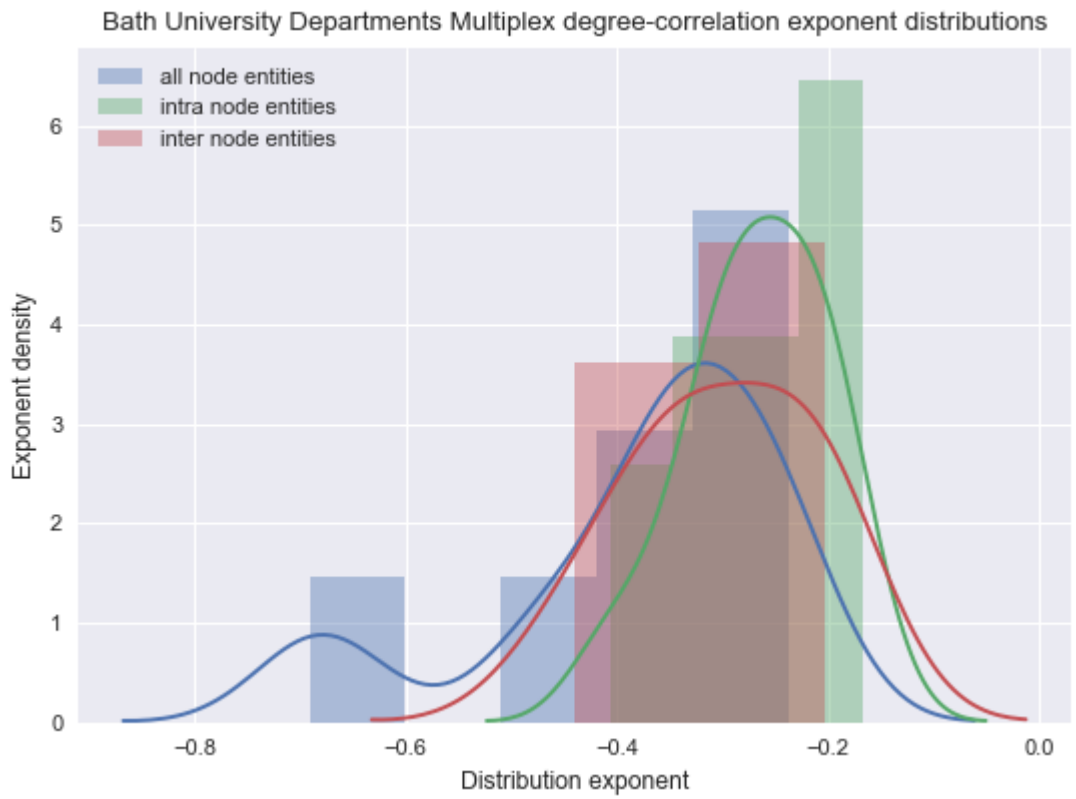

Figure 9.12. The layer degree-correlation exponent distributions for the University of Bath department-based multiplex co-authorship network. Only statistically significant exponents were included. The sample size is very small and may not be representative, but it appears to be a Gaussian distribution skewed right.

The degree correlation provides an indication as to the overall structure. A disassortative network will indicate a hub-and-spoke type structures, which generally have less clustering. This has been found for all disciplines. This could be an indication that there are a few set schools of thought or subjects of study within every discipline. Regardless of the reason why, more easily distinguishable communities can be identified that are being led by highly connected individuals. These individuals would have a high ability to influence the academic works within the community. Policy makers and decision makers would have two options to influence IDR; they could approach the leaders of the communities and propose that efforts be made to approach an IDR problem, or they could identify IDR node entities close to the community leaders.

9.3.1.4. Node activity

The node-layer activity measures how many different layers a node is active in. This can be thought of as the multiplex equivalent of degree. As can be seen in Figure 9.13, there is a strong, statistically significant negative power-law distribution. The power-law exponent is -3.32. The overall trend agrees with previously reported node activity in other networks (Nicosia and Latora 2015). This is an important finding that corroborates the power-law nature of node activity.

This means that a hypothesis can be formed regarding this measure.

Hypothesis 9.4: The multiplex node activity exhibits a power-law distribution with a negative exponent between $-2.5 \geq \gamma \geq -3.5$.

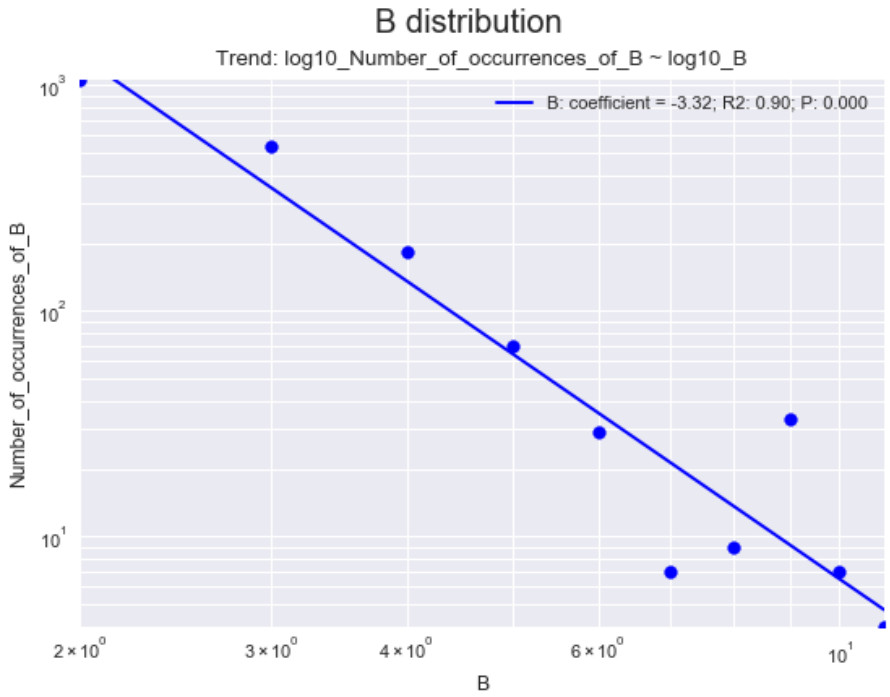

Figure 9.13. Node-layer activity of the University of Bath department-based multiplex co-authorship network. It is statistically significant and has a strong correlation with an R^2-value of 0.90.

It is worth noting that there is an outlier at $B_i = 1$ occurring a single time.

The node activity is the closest measure available to how interdisciplinary an individual is. Therefore, decision and policy makers can identify individuals who are more interdisciplinary or who better understand the issues facing IDR. Such individuals could be directly identified by higher node activity numbers.

9.3.1.5. Layer activity

As there are few layers, the sample size is too small to create a statistically significant distribution. A negative linear correlation was found as seen in Figure 9.14. Therefore, it is not possible to formulate a hypothesis for this measure.

This measure would be interesting to see if it occurs as part of a natural process, or if it is mostly driven by University policies.

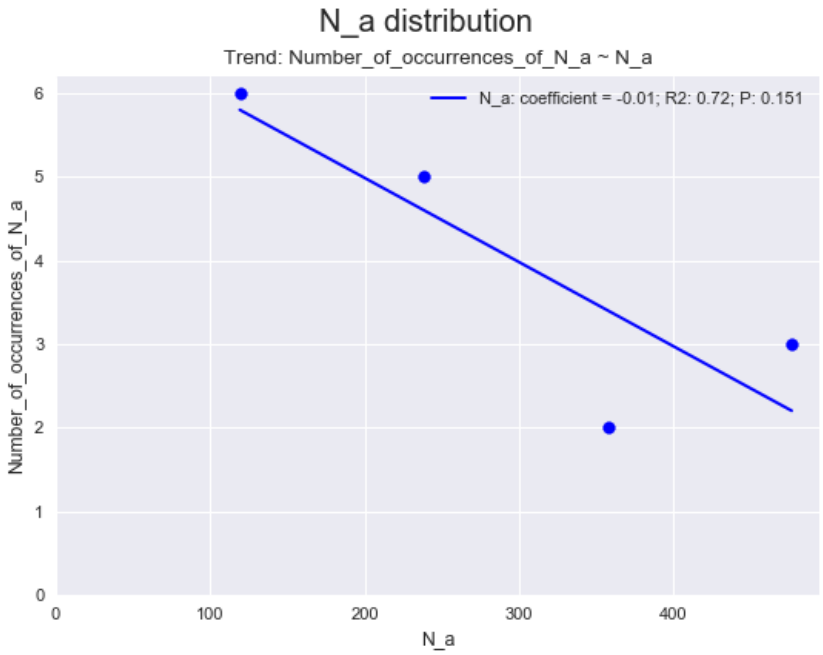

Figure 9.14 From the University of Bath department-based multiplex co-authorship network. A small sample size for layer activity results in a non-significant finding. There is still a negative linear correlation however.

The layer activity is simply the number of collaborators in every discipline. If every discipline were to be thought of as a node, it would be the weighted degree (strength). Therefore, it could be used to identify the most central discipline to the University as it also includes IDR collaborators.

9.3.1.6. Layer-pair closeness

The layer-pair closeness shows if there is a specific type of distribution for how close two specific layers are. The specific values show the number of IDR collaborators between two disciplines.

Figure 9.15 exhibits the heat map of the University of Bath interdisciplinary journal co-authors, which can provide an idea as to how IDR occurs. As can be seen, Chemistry is the most interdisciplinary discipline in the University and holds a central position amongst the departments. The question thus arises as to why Chemistry is the most interdisciplinary? It is certainly the largest department at the University of Bath, perhaps making it the discipline with the most resources? Perhaps the other disciplines' research interests are mostly centred around Chemistry? Further research would be required to answer such questions, but this effectively shows the importance of such a measure. It is able to show what disciplines drive IDR, and what disciplines work closely together.

Figure 9.15. Heatmap of interdisciplinary collaborators between disciplines 2000-2017.

As seen in Figure 9.16, it is relatively well approximated by a power-law distribution with a relatively shallow exponent agreeing with findings from Nicosia and Latora (2015).

This implies that few layer pairs have a lot of IDR occurring between them, whereas the majority of them have relatively weak IDR presences. This means that there are preferred IDR discipline pairs. Other pairs may be developing or may represent potential areas for growth, but if efficiency is a concern, established pairs can be very useful to decision-makers.

This means that a hypothesis can be formed regarding this measure.

Hypothesis 9.5: The multiplex layer-pair closeness exhibits a power-law distribution with a negative exponent between $-0.3 \geq \gamma \geq -1.0$.

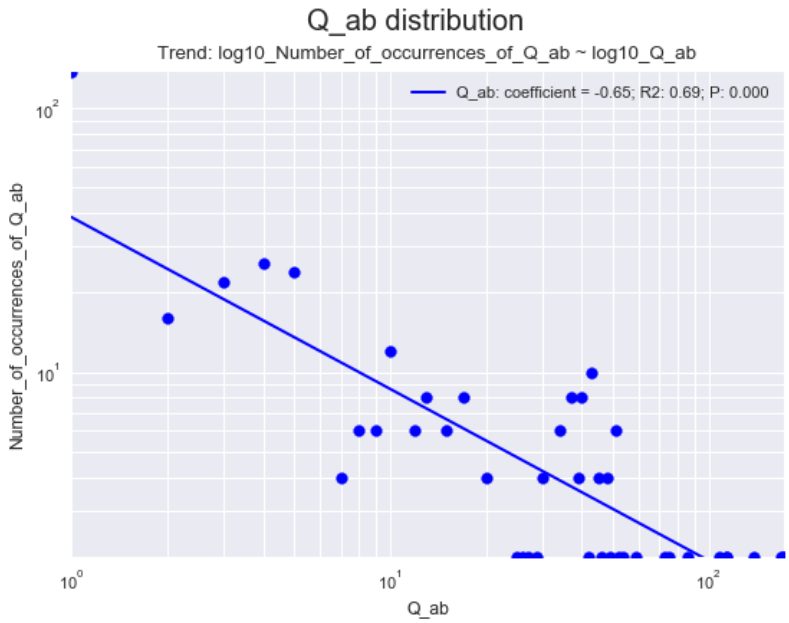

Figure 9.16. Layer closeness of the University of Bath department-based multiplex co-authorship network exhibits a power-law distribution with a shallow exponent, implying few layer pairs have a lot of IDR occurring between them, whereas the majority of them have relatively weak IDR presences.

Furthermore, when the values are summed by layers, it becomes clear that there is a negative linear trend (although it is not statistically significant as there are only 4 points), as seen in Figure 9.17. This coefficient is sharp however, and clearly shows that certain layers are much more interdisciplinary than others. As it is not statistically significant, this distribution is not codified as a hypothesis for the growth models.

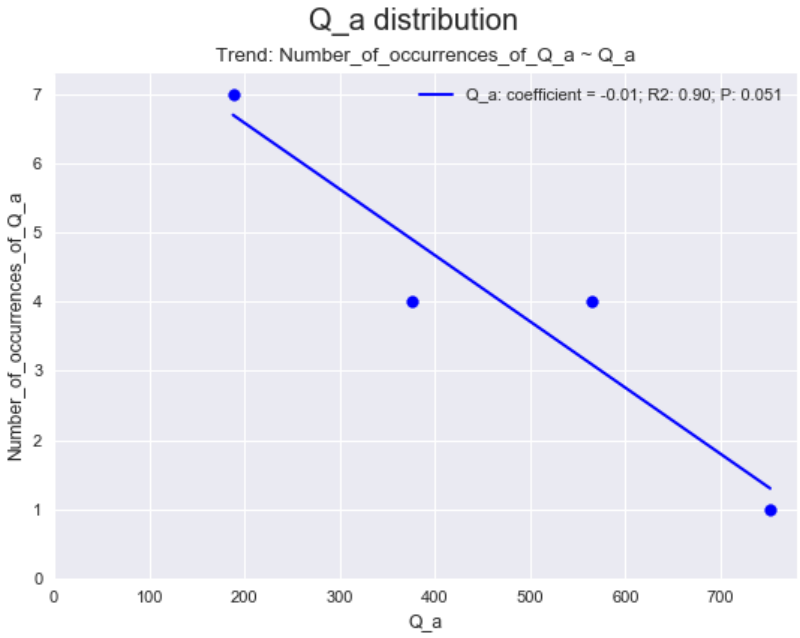

Figure 9.17. Layer closeness, of the University of Bath department-based multiplex co-authorship network, when summed by layers has a non-significant negative trend, whilst indicating certain layers are more interdisciplinary than others.

However, this implies that there are leading disciplines in IDR. As was already established, Chemistry holds that positions at the University of Bath.

From the results shown in Figure 9.17, one may be inclined to say that the layer-pair closeness represents how easily transferable knowledge is from one discipline to another (e.g. Mathematics is applicable in most disciplines). However, the reality is more complicated. Resources, core competencies, and research focus of the University itself drive such interdisciplinary research. It is likely a circular interaction occurring within the University (success in a discipline draws more funding, which in turn increases likelihood of success) that is represented in such numbers. Regardless of how it is that this particular system came to be, the current leader in IDR is Chemistry, and without a discrete change, there is no reason to believe this would change.

It also stands to reason that, with being exposed to multiple different disciplines, Chemistry has developed some expertise in conducting IDR.

As such, layer-pair closeness provides decision and policy makers with useful information from several different perspectives. It first can provide explicit information on what number of IDR collaborators exist between all different disciplines. The distribution shows that the majority of disciplines do not collaborate with each other very much. Those links between disciplines that do exist can be quite strong, and seem to be driven by a central discipline (either for its resources, position, ability to conduct IDR, or the focus that University has naturally or artificially been pushed for).

By being aware of the natural tendency to collaborate with a central field, policy or decision makers can take purposeful decisions to either support such IDR, or be aware that additional resources may be necessary to encourage other types of IDR.

9.3.2. Comparing and contrasting content-based multiplex network to the department-based multiplex network

The results have thus far described the department-based multiplex network. The content-based multiplex network exhibits very similar distributions, as shown in Table 9.5.

Table 9.5. Table comparing the trends and values of the University of Bath multiplex co-authorship networks based on department-based disciplines and content-based disciplines.

Measure		**Department-based multiplex networks**		**Content-based multiplex networks**	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-1.87	$\log_{10} y \sim \log_{10} x$	-1.87
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.55	$\log_{10} y \sim \log_{10} x$	-1.45
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-1.20	$\log_{10} y \sim \log_{10} x$	-0.70
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-2.25	$\log_{10} y \sim \log_{10} x$	-2.25
Disciplinary-interdisciplinary boxplot		$y \sim x$	0.33	$y \sim x$	0.5
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	0.05*	$\log_{10} y \sim \log_{10} x$	0.05*
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.35	$\log_{10} y \sim \log_{10} x$	-0.35**
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.28	$\log_{10} y \sim \log_{10} x$	*
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.36	$\log_{10} y \sim \log_{10} x$	*
B	**All nodes**	$\log_{10} y \sim \log_{10} x$	-3.32	$\log_{10} y \sim \log_{10} x$	-1.65**
N_a	**All nodes (only 4 points)**	$y \sim x$	-0.67*	$\log_{10} y \sim \log_{10} x$ $\log_{10} y \sim x$	-1.64* -0.00
Q_ab	**All nodes**	$\log_{10} y \sim \log_{10} x$	-0.65	$\log_{10} y \sim \log_{10} x$	-0.54
Q_a	**All nodes (only 4 points)**	$y \sim x$	-12.02*	$y \sim x$	-5.22*

*Not statistically significant.

**Significantly worse fit, lower R^2-value than its counterpart.

The majority of the exponents are very similar. For this reason, the department-based network is considered as the exemplar network. This is to reduce the complexity and repetition of the analysis. However, the similarity and differences of the two networks should be analysed in future work.

9.3.3. Discussion

The metrics defined provide a way of describing the structure of the University of Bath multiplex co-authorship network. It describes a multiplex network that contains both disciplinary and interdisciplinary node entities.

In each of the layers, a scale-free degree distribution exists. This also occurs vertically, with node activity across all layers exhibiting a similar scale-free distribution. Finally, there is a scale-free distribution for the layer-pair closeness.

These three aspects are the most descriptive of the overall structure for a multiplex co-authorship network.

Furthermore, although it has not been shown to be statistically significant, interesting dynamics are shown with regards to the degree-correlation. The aggregate degree-correlation sees a trend reversal when split into layers.

There also seems to be a negative trend with regards to the layer sizes and layer closeness centrality.

Furthermore, important differences have been found between the disciplinary and interdisciplinary node entities. As the node entities belong to the same node, it implies that there is a difference in the process, not only the person. This is a very important finding, and represents a contribution to knowledge.

The findings have provided a series of hypotheses that the growth models need to pass in order to create a realistic network. This provides a historical data validation for the growth models. The results have overall also provided a lens that the models can be compared to even though they have not been codified into hypotheses (which remains the way to determine scientific knowledge).

It is important to remember, however, that the analysis is subject to weaknesses. The analysis has chosen measures that define the overall structure of the network but does not include other established measures (e.g. clustering and the plethora of measures that have been proposed throughout literature).

9.4. Model 1: Barabási-Albert model

Having defined the criteria that the model needs to meet, it is possible to start modelling.

The Barabási-Albert model has been a standard approach to many different simulated networks (Krapivsky, Redner et al. 2000, Barabâsi, Jeong et al. 2002, Dorogovtsev and Mendes 2002). It therefore serves as strong baseline for comparison to the University of Bath multiplex co-authorship networks.

The first proposed model is a Barabási-Albert model that is grown simultaneously on all layers. There are intentionally no interdisciplinary connections. Therefore, this model is used to investigate what collaboration networks look like when no IDR occurs. This assumes that the Barabási-Albert model accurately describes disciplinary collaborations (Barabási and Pósfai 2016). However, as no IDR is being modelled, this implies that a significant component of collaboration is not being captured by the model.

Therefore, there is only a single mechanism present in this model. The Barabási-Albert growth model starts with m_0 nodes, which are connected to each other. Every timestep, another node with m_0 links is added. The links are connected to previously added nodes with a probability proportional to its degree.

$$\Phi_i^{\alpha} = \frac{k_i}{\sum_{j=1}^{N(t)} k_j} \tag{9.13}$$

Where Φ_i is the probability of a new node connecting to node *i* with degree k_i and t is the current timestep. This can be analytically shown to form a degree distribution on each layer that tends to the following degree distribution (see section 9.4.3.1.).

$$p(k) \sim k^{-3} \tag{9.14}$$

As this model does not have any interdisciplinary connections, only the degree distributions and degree-correlations are considered.

9.4.1. Degree distribution

The simulation results are given in this section, and compared to the analytical results, and the real results. This section determines whether Hypothesis 9.1 is corroborated.

Hypothesis 9.1: The degree distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.

The aggregate degree distribution for the simulation is given in Figure 9.18. The exponent is given as -2.76, which is relatively close to the -3 that it tends to analytically.

However, in terms testing Hypothesis 9.1(a), this exponent is too large. Hypothesis 9.1(a) is therefore rejected.

Hypothesis 9.1(a) - The aggregate degree distribution produces a power-law relationship with an exponent between -1.5 to -2.5

Therefore, a pure Barabási-Albert model is not a good representation, and other mechanisms affecting the distribution are needed to reduce the exponent.

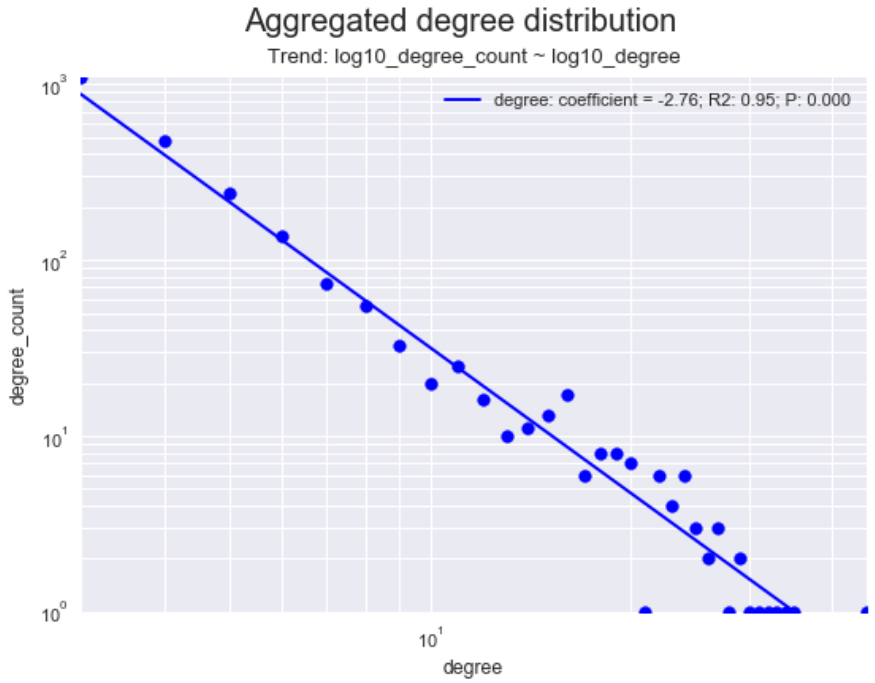

Figure 9.18. The aggregate degree distribution for the Barabási-Albert model grown on individual layers with no overlap.

When inspecting the individual layers in Figure 9.19, the exponents are much smaller. As the distribution requires fewer points at larger degrees (smaller number of occurrences), there is scope for these high degree points to skew the distribution. This may in part explain why the OLS trend does not fit the line the points seem to make. Furthermore, the Barabási-Albert model analytically tends towards a power-law distribution with an exponent of -3.

However, for the simulation parameters used, Hypothesis 9.1(b) passes.

Hypothesis 9.1(b) - The degree distribution on every layer produces a power-law relationship using all node entities, disciplinary node entities only, and interdisciplinary node entities only.

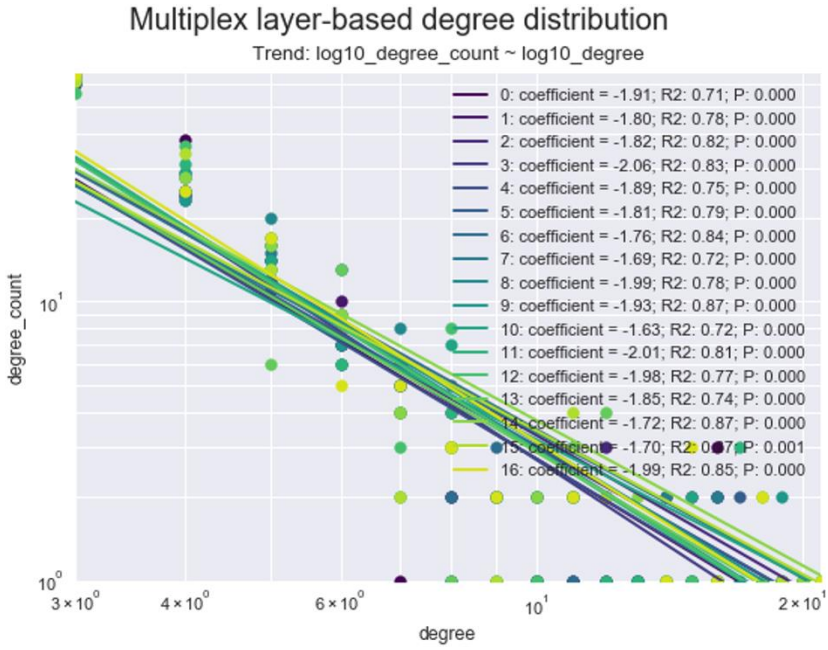

Figure 9.19. Layer degree distribution for all node entities model 1: Barabási-Albert model.

The layers' exponents produce a Gaussian distribution as shown in Figure 9.20. The exponents are all below the aggregate's value, thereby Hypothesis 9.1(c) passes.

Hypothesis 9.1(c) - The degree distribution on every layer, using every node entity, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than the aggregate exponent.

However, it forms a Gaussian distribution without any skewness. Thereby rejecting Hypothesis 9.1(f).

Hypothesis 9.1(f) - The degree distributions' exponents are distributed as Gaussians that are skewed to the right as estimated by the KDE.

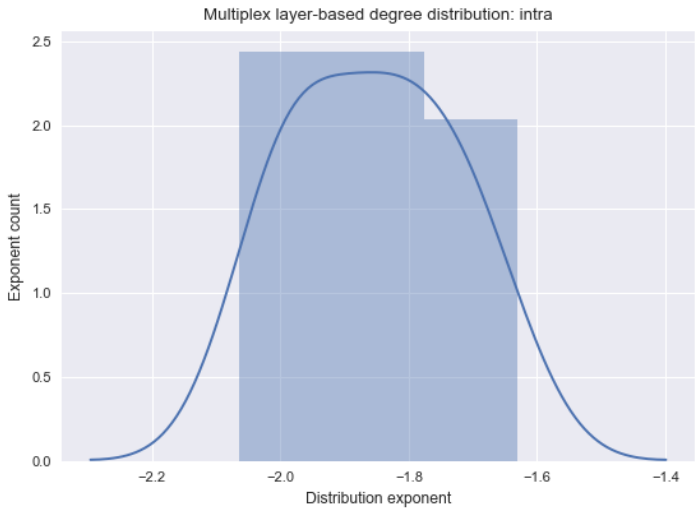

Figure 9.20. Exponent distribution for model 1: Barabási-Albert algorithm.

Of the possible hypotheses that could be passed, the model rejected Hypotheses 9.1(a) and 9.1(f). This suggests that degree distribution is partially valid but requires changes to make it truly valid including reducing the degree distributions' exponents (magnitude) and skewing the layers' exponents distribution to the right.

9.4.2. Degree-correlations

The degree-correlations are expected to match Barabási-Albert analytical analysis, which would indicate that the degree-correlation is very small and negative (see Analytical analysis). This matches the real-world degree-correlation.

> ***Hypothesis 9.3: The degree-correlation distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.***

The simulation results show a clear negative trend in both the aggregate and layer degree-correlations as shown in Figure 9.21 and Figure 9.22 respectively. It is worth noting that the exponents are significantly smaller than analytically predicted as shown in Figure 9.23. This is most likely due to the several approximations that have been assumed throughout the analysis.

The corresponding real-world degree-correlation was statistically insignificant to the 0.05 significance level. However, it was significant to the 0.1 significance level and exhibited a positive power-law exponent. This does not match to the model, and opposite trends are seen (significant to the 0.1 significance level). This does not reject any hypothesis.

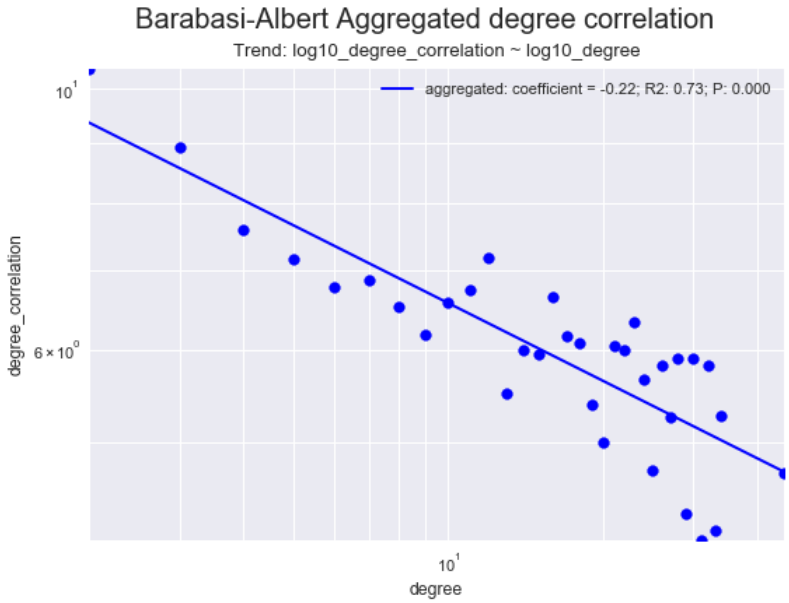

Figure 9.21. The degree-correlation for the aggregate network created in model 1: Barabási-Albert. A significant negative trend is found.

However, each layer does exhibit a power-law distribution with a negative exponent as shown in Figure 9.22. Hypothesis 9.3(a) therefore passes.

Hypothesis 9.3(a) - Layers exhibit degree-correlation distributions with a power-law relationship with a negative exponent.

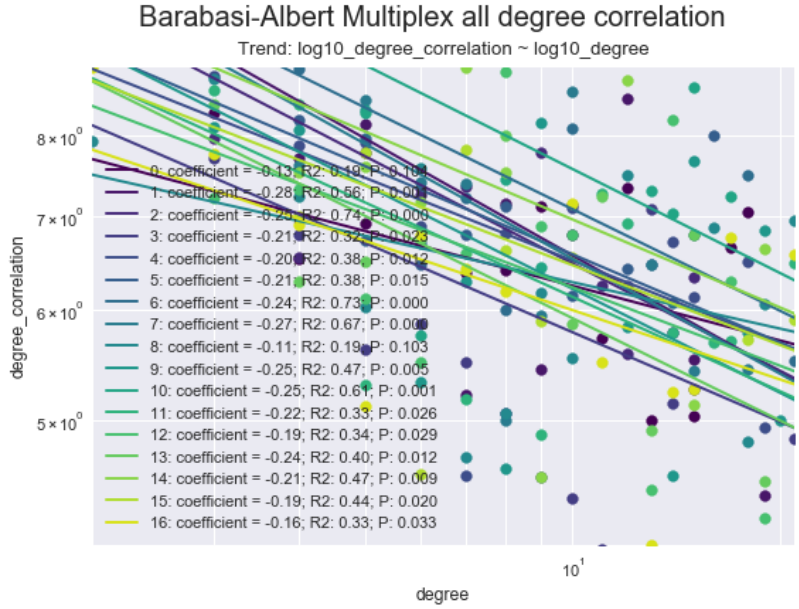

Figure 9.22. The degree-correlation for the network layers created in model 1: Barabási-Albert. Each layer exhibits a power-law distribution with a negative exponent.

Examining the degree-correlation exponent distribution in Figure 9.23, Hypothesis 9.3(b) is partially corroborated, but is not skewed to the right.

Hypothesis 9.3(b) - Degree-correlation distribution exponents exhibit Gaussian distributions as estimated by the KDE skewed right with the peak density occurring at a value of $\gamma > -0.3$.

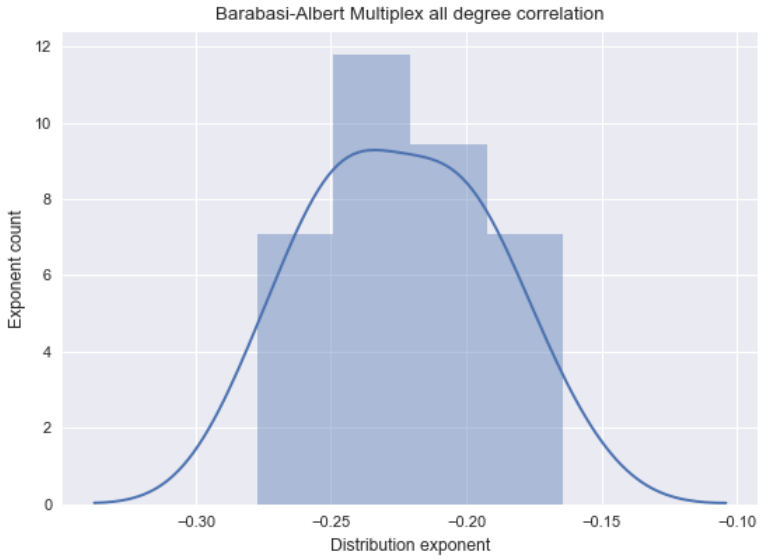

Figure 9.23. The degree-correlation distributions of the Barabási-Albert model network's exponents. Gaussian distribution occurs, but the right skew is lacking.

9.4.3. Analytical analysis

Having analysed the model results, it is useful to understand how it is that the model affects properties of interest. This section outlines the analytical analysis of the degree distribution and the degree correlation for the Barabási-Albert model. Analytical analyses can be used to examine how the various structural properties progress with time, which can provide great insight on factors and variables of importance. In a pure Barabási-Albert algorithm, as there are only two possible factors to consider: the degree of a node and the size of a network. The reason that the analytical analysis is so important is that the Barabási-Albert model approximates the probability of a node degree to be proportional to k^{-3} and not at all to the size of the network, ergo most real networks have since been called 'scale-free' (as they do not scale with the size of the network).

The analytical analyses in this research provides a means to draw similar type of conclusions about the nature of multiplex collaboration networks.

9.4.3.1. Degree distribution analytical solution

Following the analysis as outlined in Barabási and Pósfai (2016), the following expression can be approximated for Barabási-Albert algorithm (see Appendix B for the derivation).

$$p(k) = 2{m_0}^2 k^{-3} \tag{9.15}$$

As m_0 is simply a constant, the degree distribution follows the power-law exponent -3. This provides an approximate measure, an exact solution is provided in Appendix, but amounts to the same conclusion.

For the aggregate network, when growing the multiplex network layer-by-layer, the dynamics changes a little. This analysis is absolutely valid for individual layers, if each layer is treated as a separate network.

$$p^{\alpha}(k) = 2{m_0}^2 k^{-3} \tag{9.16}$$

The only difference occurs when trying to find the aggregate degree distribution. The aggregate distribution can easily be found.

$$p^{\forall\alpha}(k) = \frac{1}{N}\sum_{\alpha=1}^{M} p^{\alpha}(k) \cdot N^{\alpha} \tag{9.17}$$

Where $N^{\alpha} = m_0 + t$ and $N = (m_0 + t)M$

$$p^{\forall\alpha}(k) = \frac{1}{(m_0 + t)M}\sum_{\alpha=1}^{M} 2{m_0}^2 k^{-3} \cdot (m_0 + t) \tag{9.18}$$

$$p^{\forall\alpha}(k) = 2{m_0}^2 k^{-3} \tag{9.19}$$

The same power-law will still hold, $p^{\forall\alpha}(k) \sim k^{-3}$ for $k \gg m_0$.

This means that the degree distribution that has been experienced in most real networks, can be approximated with a Barabási-Albert algorithm (Newman 2010, Barabási and Pósfai 2016). The algorithm always tends towards a power-law exponent of -3, regardless of network size, ergo such networks are described as scale-free (the structure remains the same regardless of size). This analysis shows that the scale-free property does not change regardless of how many layers are grown simultaneously.

Therefore, extending growth models to be grown individually on all layers is a suitable approach to approximating multiplex networks. It also implies that the many findings surrounding Networks Science can be applied on each layer and the network as a whole.

9.4.3.2. Degree-correlation distribution analytical solution

The degree-correlation analysis starts by establishing the degree-correlation of a node. It is given by the following expression.

$$k_{nn}(k_i) = \frac{\sum_{j=1}^{N} A_{ij} k_j}{k_i} \tag{9.20}$$

An analytical solution suggests that a Barabási-Albert algorithm will yield a neutral degree-correlation (Barrat and Pastor-Satorras 2005). However, they all require some asymptotic approximation of $\sum_{j=1}^{N} A_{ij} k_j$, which becomes an approximation of k_i.

Barrat and Pastor-Satorras (2005) find that the Barabási-Albert algorithm can be approximated by the following expression.

$$k_{nn}(k) = \frac{m_0}{2} \ln(N) \tag{9.21}$$

This suggests that the Barabási-Albert algorithm is not dependent on the degree, and is therefore a neutral degree-correlation regime. This approach was achieved using an Ordinary Differential Equation.

It is possible to write the equation in terms of t_i, t_j, and t and get a different solution by integrating the rate of change of the numerator.

$$\sum_{j=1}^{N} A_{ij} k_j(t+1) = \sum_{j=1}^{N} A_{ij} k_j(t) + m_0 \frac{dk_i}{dt} + \sum_{j=1}^{N} A_{ij} \frac{dk_j}{dt} \tag{9.22}$$

This can be thought of as the future sum of all neighbours' (of node *i*) degrees is equal to the current sum of all neighbours' degrees, plus the rate of change the degree of node *i* multiplied by the degree of new nodes (m_0), plus the rate of change of all current neighbours' degrees.

Plugging in the expressions for $\frac{dk_i}{dt}$, $\frac{dk_j}{dt}$, k_i, and k_j.

$$\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t+1}{t_j}\right)^{\frac{1}{2}}\right)=\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t}{t_j}\right)^{\frac{1}{2}}\right)+\frac{\left(m_0\left(m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}\right)+\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t}{t_j}\right)^{\frac{1}{2}}\right)\right)}{2m_0t+(m_0^2-m_0)} \tag{9.23}$$

$$\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t+1}{t_j}\right)^{\frac{1}{2}}\right)=\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t}{t_j}\right)^{\frac{1}{2}}\right)\left(\frac{2m_0t+(m_0^2-m_0)+1}{2m_0t+(m_0^2-m_0)}\right)+\frac{m_0\left(m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}\right)}{2m_0t+(m_0^2-m_0)} \tag{9.24}$$

As $t \gg m_0 \geq 1$ at later timesteps.

$$\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t+1}{t_j}\right)^{\frac{1}{2}}\right)=\sum_{j=1}^{N} A_{ij}\left(m_0\left(\frac{t}{t_j}\right)^{\frac{1}{2}}\right)+\frac{m_0\left(m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}\right)}{2m_0t} \tag{9.25}$$

$$\sum_{j=1}^{N} A_{ij}k_j(t+1)-\sum_{j=1}^{N} A_{ij}k_j(t)=\frac{m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}}{2t} \tag{9.26}$$

Using a forward time-stepping scheme, it is possible to express the different time-steps as a single derivative

$$\frac{d\left(\sum_{j=1}^{N} A_{ij}k_j(t)\right)}{dt}=\frac{\sum_{j=1}^{N} A_{ij}k_j(t+1)-\sum_{j=1}^{N} A_{ij}k_j(t)}{\Delta t}=\frac{m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}}{2t\Delta t} \tag{9.27}$$

Where $\Delta t = 1$ in this simulation without exception.

Integrating this expression.

$$\sum_{=1}^{N} A_{ij}k_j(t)=\frac{1}{2}m_0\frac{1}{t_i^{\frac{1}{2}}}\int t^{-\frac{1}{2}}\cdot dt \tag{9.28}$$

$$\sum_{j=1}^{N} A_{ij}k_j(t)=\frac{1}{2}m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}}+C \tag{9.29}$$

Applying boundary condition to find the constant.

$$\sum_{j=1}^{N} A_{ij}k_j(t_i)=\frac{1}{2}m_0\left(\frac{t_i}{t_i}\right)^{\frac{1}{2}}+C=m_0\langle k\rangle=m_0\ln(N) \tag{9.30}$$

$$C = \cancel{\frac{1}{2}m_0} + m_0\ln(N) \cancel{- \frac{1}{2}m_0} \tag{9.31}$$

Applying the constant in the original expression.

$$\sum_{j=1}^{N} A_{ij}k_j(t) = \frac{1}{2}m_0\left(\frac{t}{t_i}\right)^{\frac{1}{2}} + m_0\ln(N) \tag{9.32}$$

Dividing by k_i provides the degree-correlation.

$$k_{nn}(k_i) = \frac{\sum_{j=1}^{N} A_{ij}k_j}{k_i} = \frac{1}{2} + \frac{m_0\ln(N)}{k_i} \tag{9.33}$$

This analysis was conducted to consolidate consistent differences seen in the Barabási-Albert model and expressions reported in literature (Barrat and Pastor-Satorras 2005). The Barabási-Albert model has reported a neutral degree It has been reported that a neutral degree correlation is expected, whereas there is consistently a power-law with a negative exponent visible in the simulations.

The alternative approach presented here suggests that a negative exponent is possible.

Contribution to knowledge

The negative degree correlation value found for the University of Bath 2000-2017 disciplines and the Barabási-Albert algorithm prompted a re-examination of the algorithm's analytical analysis as it did not seem to be neutral. By using forward time-stepping scheme, an alternative analysis is approximated to the following equation.

$$k_{nn}(k) \sim \frac{1}{2} + m_0 \ln(N) k^{-1} \tag{9.34}$$

This analysis does not agree with the generalised solution that Barrat and Pastor-Satorras (2005) found for this implementation of the Barabási-Albert algorithm. This approach predicts a negative power-law correlation, which matches the results (albeit results suggest an exponent of -0.22, not -1). This represents a minor original contribution to knowledge. The actual results from the algorithm seem to fall somewhere between the two approaches, which is likely explained by the assumptions made in the respective analyses.

9.4.4. Discussion

In comparison to real results, it can clearly be seen that this model is an overall poor fit. This is simply due to the fact that the majority of the hypotheses are not even testable as there are no interdisciplinary node entities or links.

However, of those that were testable, the results were reasonable considering how limited the model is. Other than the layer exponents for both the degree and degree-correlations being skewed to the right in the real network and the degree exponents being too high, a reasonable starting point has been achieved.

A few interesting similarities were identified. For instance, the aggregate degree exponent being larger than its layers' exponents. This is likely a result of the mechanics of the mulitplex network, and may always pass. The layer degree-correlations are also the correct order of magnitude, suggesting that a neutrally correlated network is a decent approximation.

Table 9.6. A comparison of the Barabási-Albert simultaneous growth model to the University of Bath department-based multiplex co-authorship network.

Measure		**Department-based multiplex networks**		**Barabási-Albert algorithm grown simultaneously on layers**	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-1.87	$\log_{10} y \sim \log_{10} x$	-2.76
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.55; Skewed right	$\log_{10} y \sim \log_{10} x$	-1.9; Gaussian
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	0.05*	$\log_{10} y \sim \log_{10} x$	-0.22
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.33; Skewed right	$\log_{10} y \sim \log_{10} x$	-0.23; Gaussian

Ultimately, this information provides no actionable information. It only provides information for modellers. It establishes that for the 'horizontal' structure (i.e. within a layer), the Barabási-Albert algorithm provides a reasonable approximation. It also shows that the phenomenon that aggregate networks exhibit larger exponents is somewhat inherent, but without overlapping presences (i.e. IDR), the exponent is far too large.

Therefore, the Barabási-Albert model is a reasonable starting point, but is missing a vital IDR component.

9.5. Model 2: Barabási-Albert model with randomly assigned layers

The first model established that the Barabási-Albert model provided a reasonable approximation given that no IDR was included. The second model establishes what a model would look like if there were no barriers to IDR at all. The fundamental assumption here is that if there are no barriers to IDR, then disciplines would play no role in how collaboration occurs. For that reason, the model creates a single layer Barabási-Albert model, then randomly assigns core disciplines to every node, and then splits up the network into layers (creating node entities in every layer in the process).

This achieves a multiplex network that establishes what a network would look like if there are no barriers between the different layers.

This will undoubtedly create a lot of overlap, as every node-neighbour pair only has $\frac{1}{M}$ chance of being neighbours on the same core-discipline, which is unrealistic. This is, however, the purpose of this model, to show what is unrealistic, and what interdisciplinary nodes would look like if there were no barriers to performing IDR.

9.5.1. Degree distribution

This section establishes whether the model's degree distribution matches the real network's degree distribution.

Hypothesis 9.1: The degree distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.

As with the previous model, this is an exact Barabási-Albert algorithm implementation on the aggregate layer and has been shown analytically to tend towards a power-law relationship with an exponent of -3.

The aggregate degree distribution is shown in Figure 9.24, and exhibits an OLS exponent of -2.06. This fit seems to be relatively poor, and it appears that a larger magnitude would be a better fit. However, with the OLS prediction, Hypothesis 9.1(a) is corroborated.

Hypothesis 9.1(a) - The aggregate degree distribution produces a power-law relationship with an exponent between -1.5 to -2.5

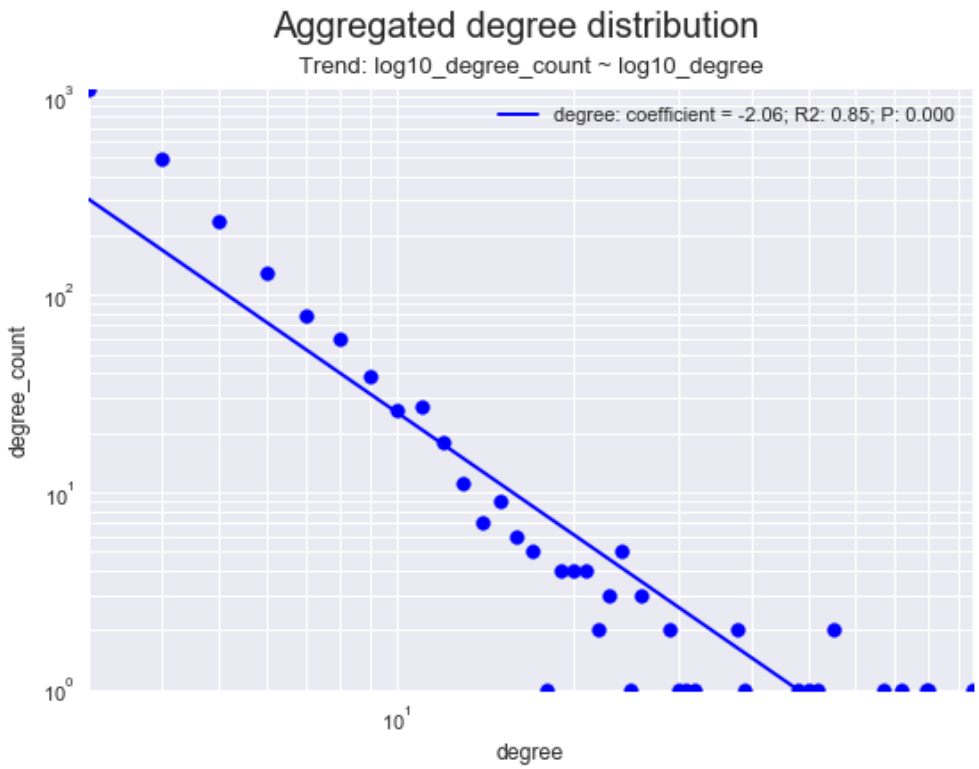

Figure 9.24. The aggregate degree distribution for the Barabási-Albert model grown and then split into layers assigned randomly.

The individual layers form power-law distributions across all node entities, disciplinary node entities, and interdisciplinary node entities as can be seen in Figure 9.25 to Figure 9.27. This means that Hypothesis 9.1(b) is corroborated.

> ***Hypothesis 9.1(b) - The degree distribution on every layer produces a power-law relationship using all node entities, disciplinary node entities only, and interdisciplinary node entities only.***

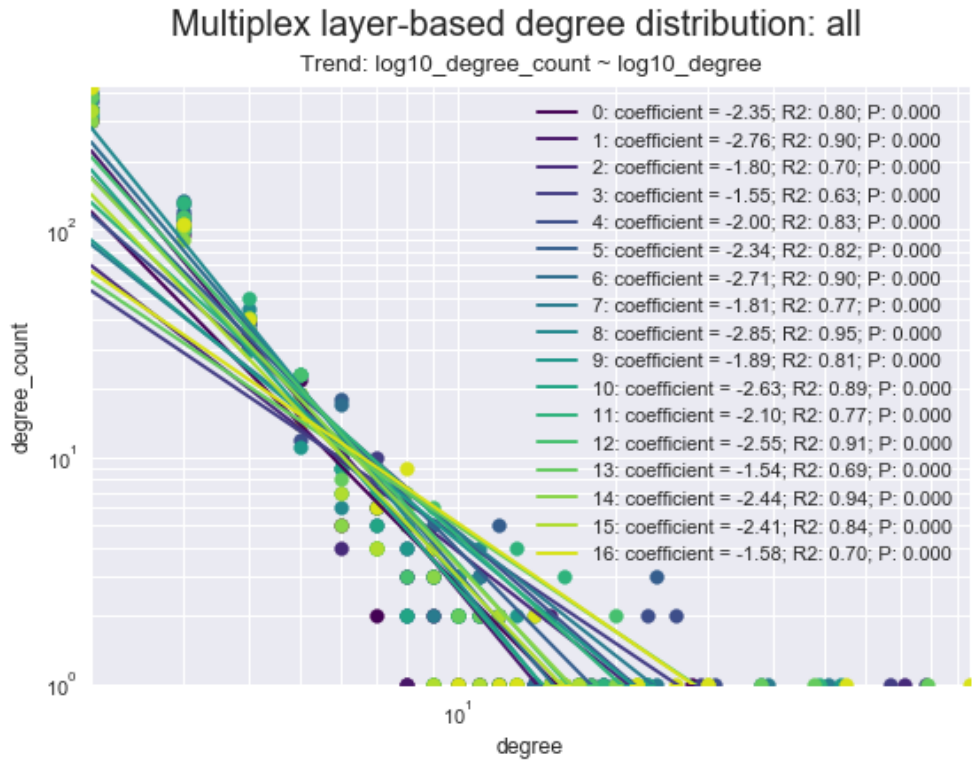

Figure 9.25. Layer degree distribution for all nodes for the Barabási-Albert model split into randomly assigned layers.

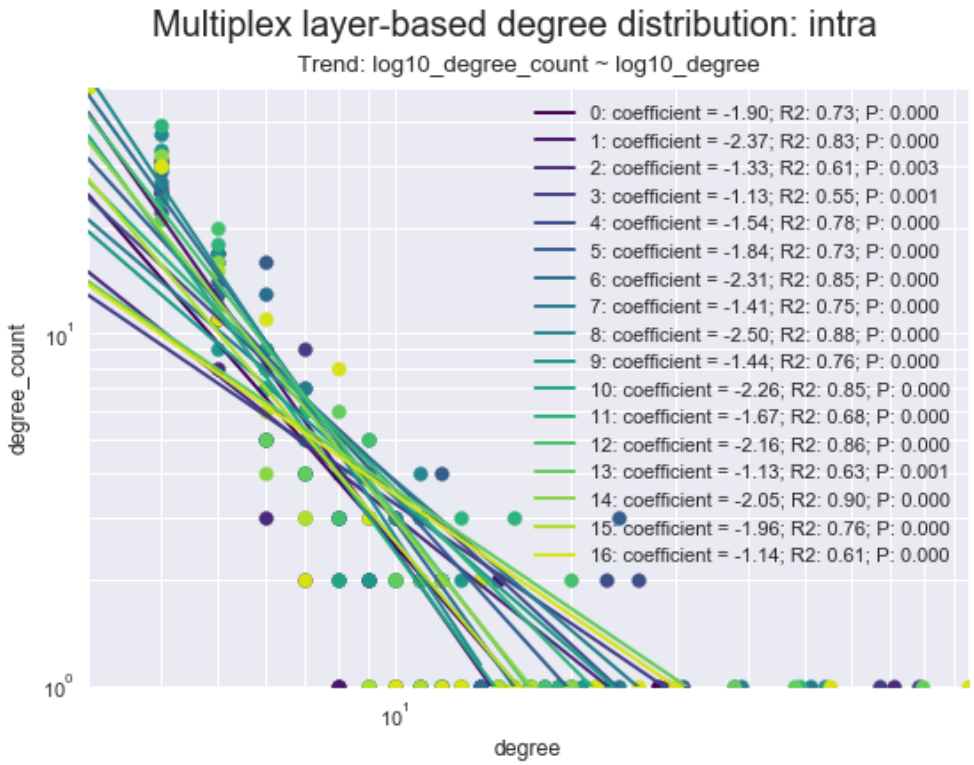

Figure 9.26. Layer degree distribution for disciplinary nodes only for Barabási-Albert model split into randomly assigned layers.

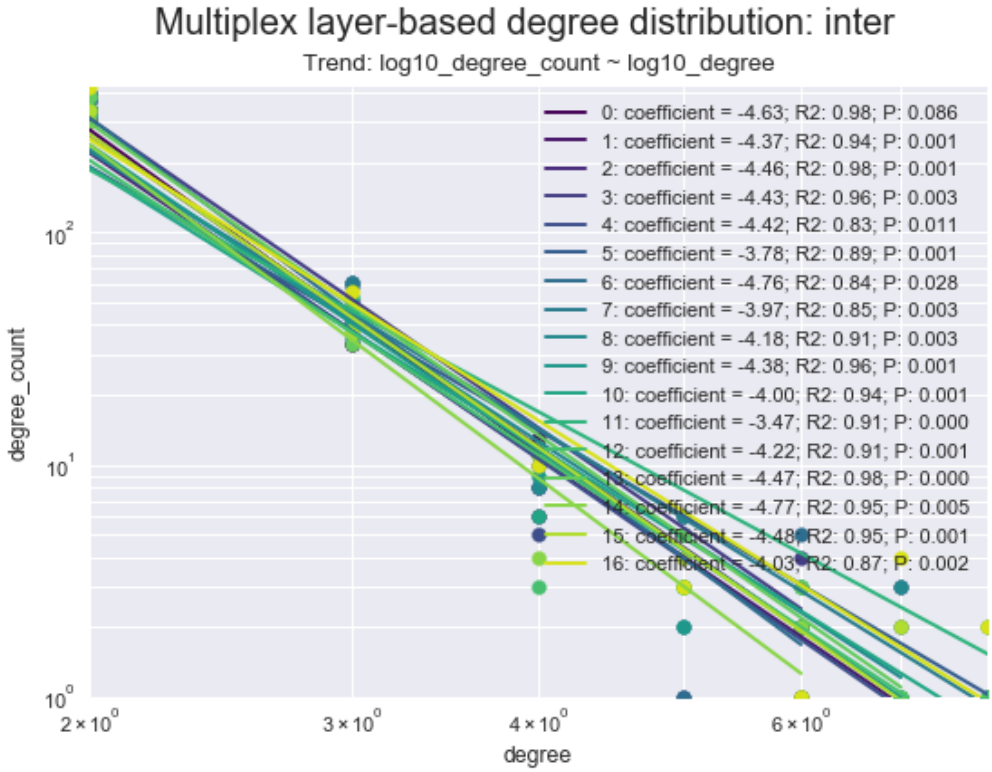

Figure 9.27. Layer degree distribution for interdisciplinary nodes only for Barabási-Albert model split into randomly assigned layers.

Figure 9.28 shows the distribution of exponents. As can clearly be seen, Hypotheses 9.1(c)-(f) are corroborated. The degree distribution when randomly assigned layers creates a Gaussian distribution that is skewed right. However, the magnitude of the interdisciplinary node entities' exponents are far too large.

Hypothesis 9.1(c) - The degree distribution on every layer, using every node entity, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than the aggregate exponent.

Hypothesis 9.1(d) - The degree distribution on every layer, using disciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than all the node entities' peak exponent.

Hypothesis 9.1(e) - The degree distribution on every layer, using interdisciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent above the aggregate exponent.

Hypothesis 9.1(f) - The degree distributions' exponents are distributed as Gaussians that are skewed to the right as estimated by the KDE.

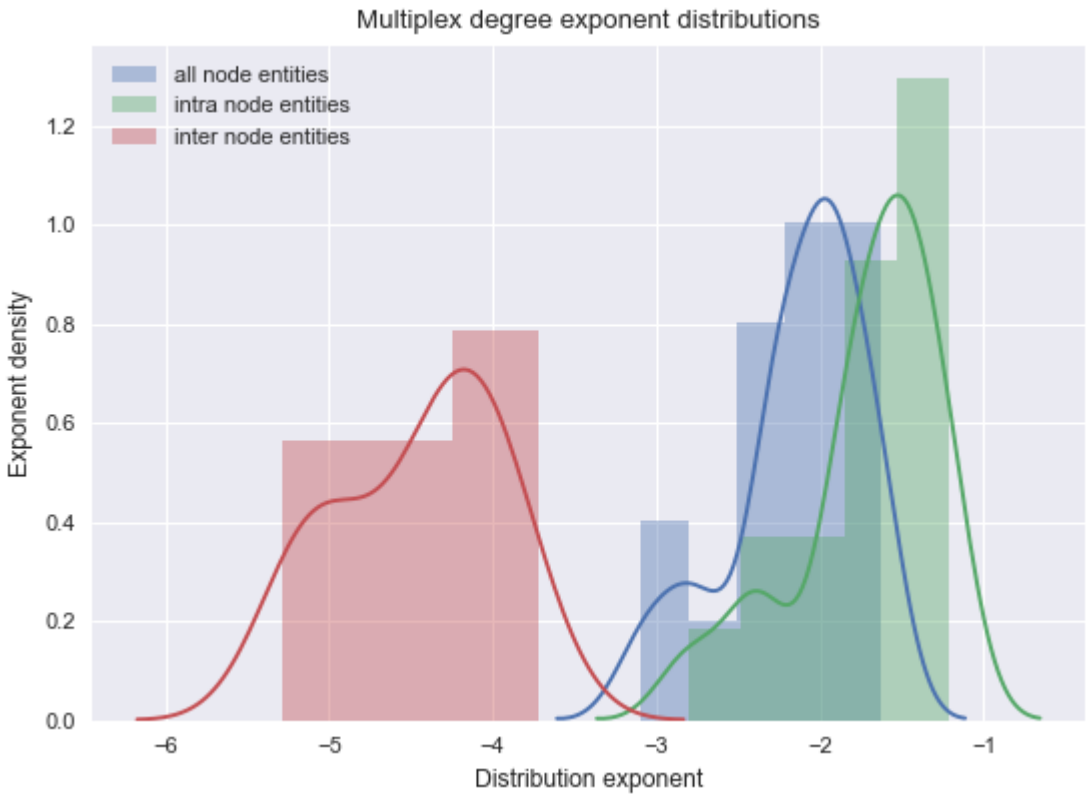

Figure 9.28. Exponent distribution for the Barabási-Albert algorithm.

It is difficult to explain why the exponents skew right when normal Barabási-Albert models show no skewness. It must therefore occur due to the way the layers as split. As there is equal probability for all layers to have all nodes, it must be driven by nodes with higher degrees. These nodes will have a multiplex node activity proportional to its degree, thereby reducing the exponent (gradually smoothing out the fat-tail), skewing the distributions right.

Ultimately, the aggregate degree distribution will be representative of the Barabási-Albert algorithm. The layer specific distributions are random samples of the network due to the existence of node entities.

As core disciplines are randomly assigned, IDR node entities will be far more numerous in a specific layer, but be far more limited to the number of node entities they can connect with (i.e. disciplinary

node entities are far more likely to connect to IDR node entities and IDR node entities generally have small degrees, but are numerous). This is not seen in real networks.

9.5.2. Disciplinary vs interdisciplinary degree regression

Comparing the disciplinary node entities' degrees to the sum of their interdisciplinary counterparts produces a plot that is about as expected. The random probability implies that there is no barrier or preference to being disciplinary or interdisciplinary, and by virtue of many layers existing, a link only has a $\frac{1}{M}$ chance of being disciplinary.

Whilst it is predictable, it is important to compare this result to the University of Bath multiplex co-authorship networks' results. The real network shows a clear preference for disciplinary links, as it stays below the $k_{intra} = k_{inter}$ line.

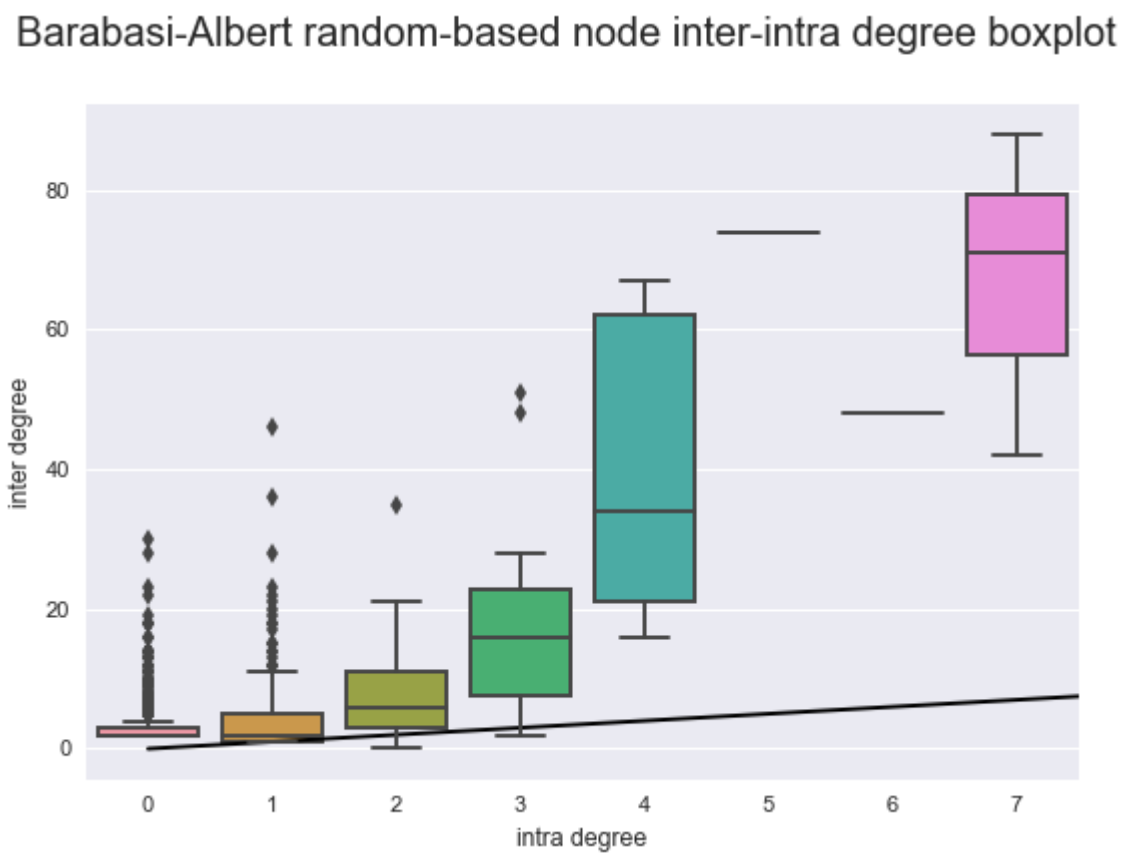

Figure 9.29. Disciplinary node entities degree vs. the interdisciplinary node entities' sum of degrees for Model 2.

This measure shows the problem with randomly assigning layers. As there are far more nodes outside any one layer, the proportion of interdisciplinary links is unrealistic. Hypothesis 9.2 is resoundingly rejected.

Hypothesis 9.2: The disciplinary node entities degrees are larger than the median of the sum of their counterpart interdisciplinary node entities' degrees.

It is worth noting that there seems to be a non-linear element. This occurs because highly connected nodes will have a large presence in many other layers. There is therefore a natural tendency for highly connected nodes to being more interdisciplinary. As this is not observed in real networks, it implies that highly connected nodes stay within certain disciplines rather than their interdisciplinarity being proportional to their degree.

The most important aspect to note is that this is what the graph would look like in a real network if there were no barriers to conducting IDR (i.e. everyone collaborated with disciplines randomly, or if disciplines did not exist).

9.5.3. Degree-correlations

This section establishes whether the degree-correlations in this model match the degree-correlations of the real-world network.

Hypothesis 9.3: The degree-correlation distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.

The degree-correlations behave as a normal Barabási-Albert algorithm would. There is a very slight negative power-law trend on all aggregations as can be seen in Figure 9.30, Figure 9.31, Figure 9.32, and Figure 9.33. Thereby corroborating Hypothesis 9.3(a)

Hypothesis 9.3(a) - Layers exhibit degree-correlation distributions with a power-law relationship with a negative exponent.

Figure 9.34 exhibits the exponents distributions and show that the degree-correlation for the layers go significantly too negative. This is driven by the interdisciplinary node entities, as the disciplinary node entities exhibit reasonable exponents. Therefore, Hypothesis 9.3(b) is rejected.

Hypothesis 9.3(b) - Degree-correlation distribution exponents exhibit Gaussian distributions as estimated by the KDE skewed

right with the peak density occurring at a value of $\gamma >$ 0.3.

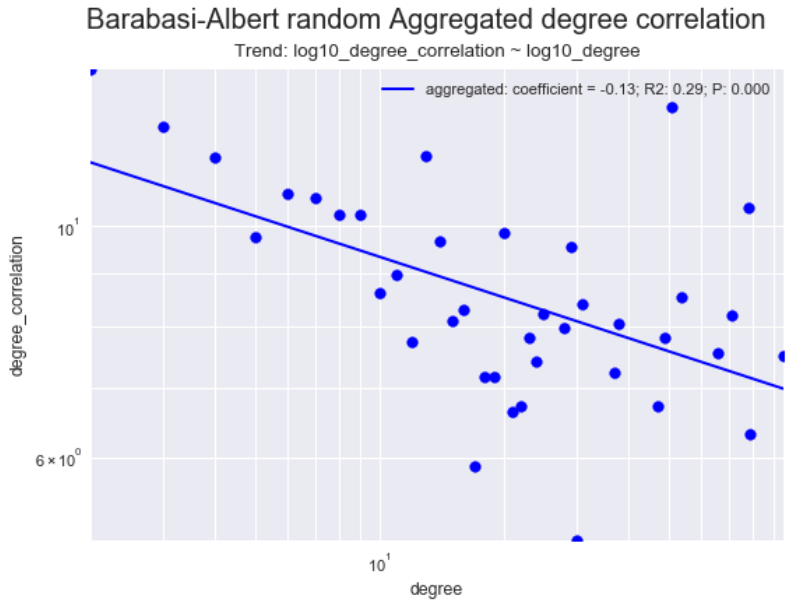

Figure 9.30. The degree-correlation for the Barabási-Albert model aggregate network. Whilst it is significant, there is poor correlation, and is negative.

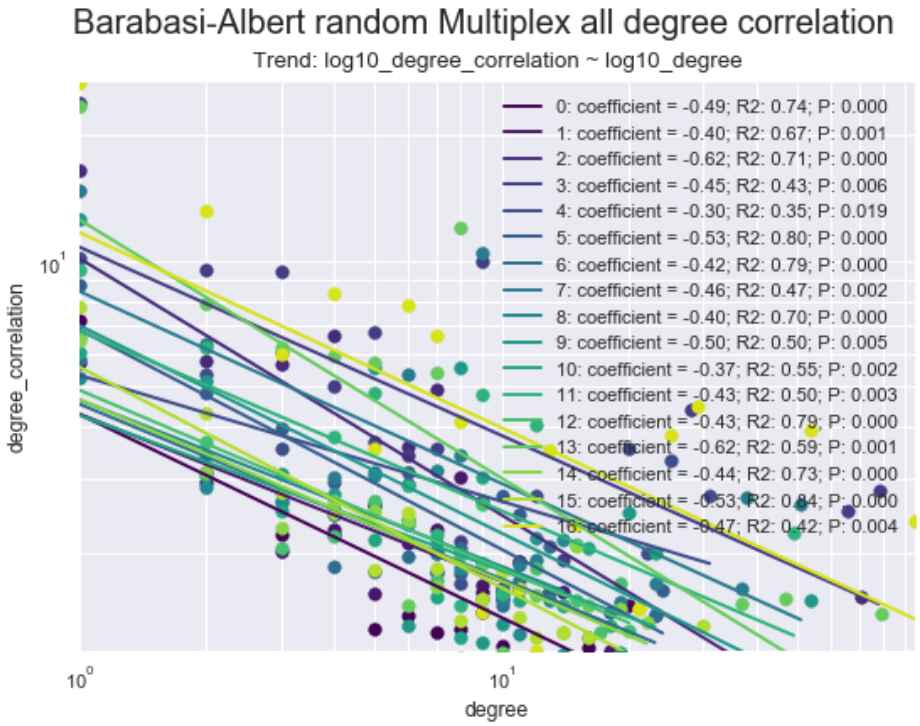

Figure 9.31. The degree-correlation for the Barabási-Albert model for all nodes.

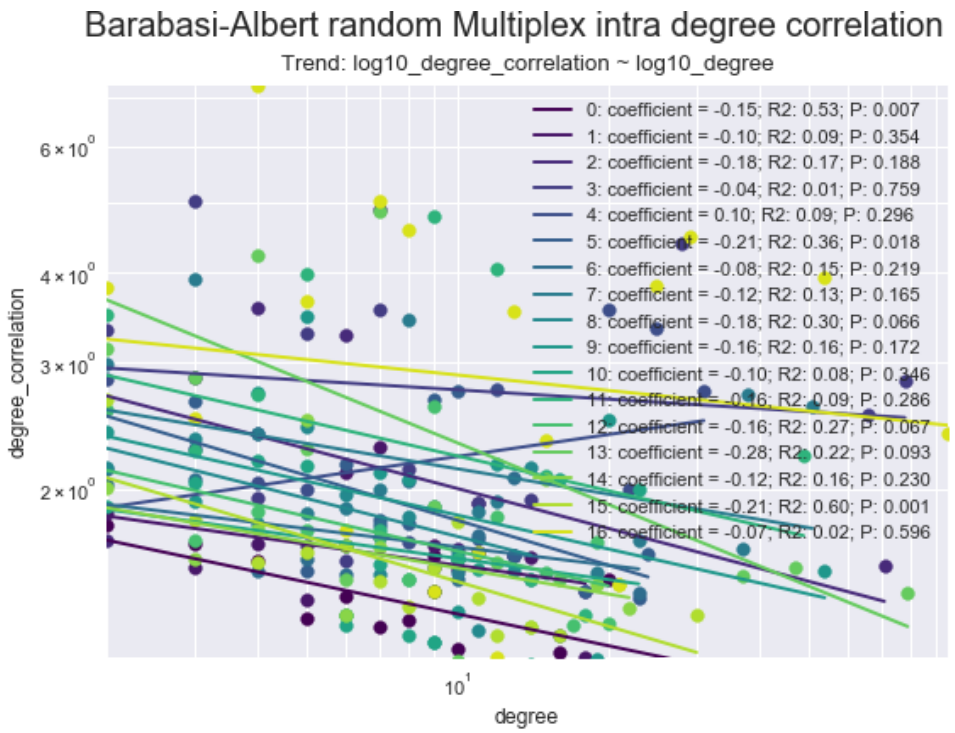

Figure 9.32. The degree-correlation for the Barabási-Albert model for disciplinary nodes only.

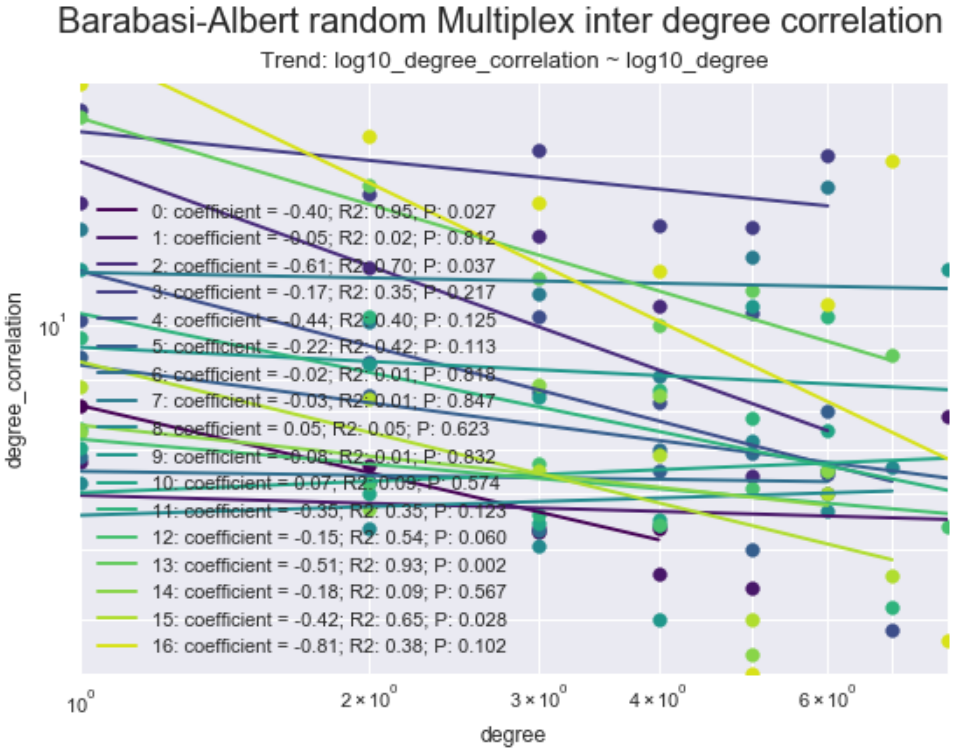

Figure 9.33. The degree-correlation for the Barabási-Albert model for interdisciplinary nodes only.

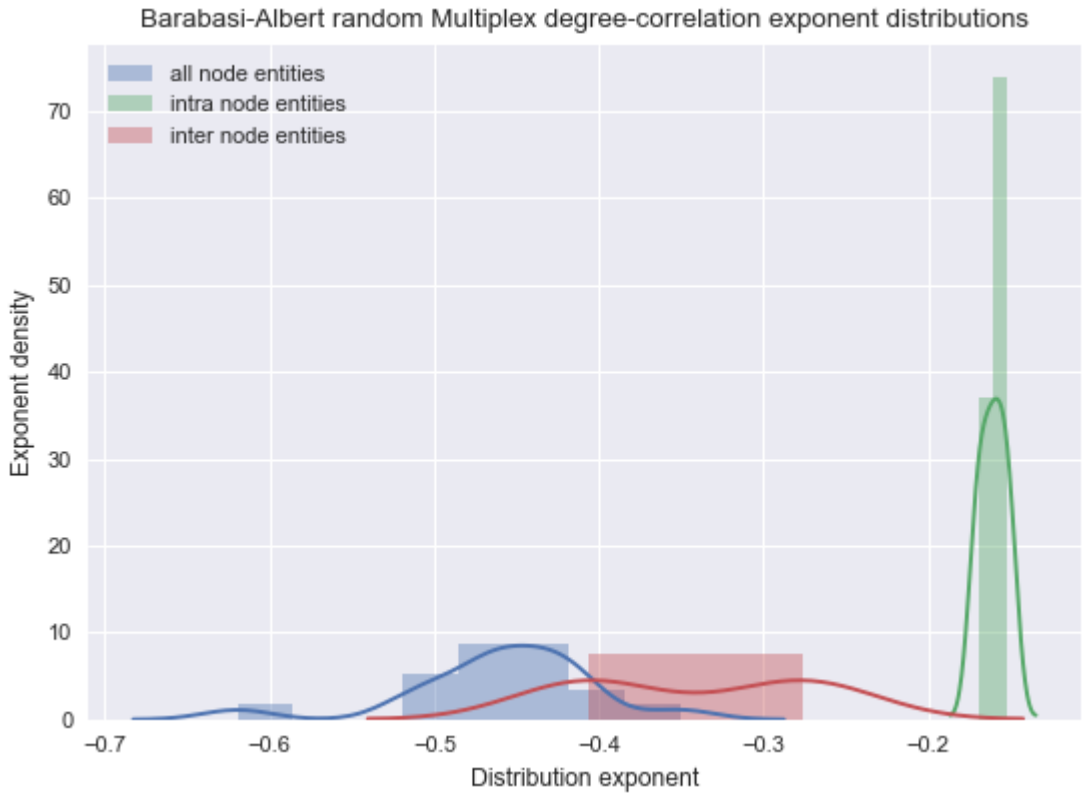

Figure 9.34. The layer degree-correlation distributions of the Barabási-Albert model network's exponents. Degree-correlation is significantly too negative.

9.5.4. Node activity

The node-layer activity produces a fat-tailed Poisson distribution as shown in Figure 9.35. This is in stark contrast to the University of Bath multiplex co-authorship network, which produces a clear power-law with exponents of -3.32 and -1.65 for the department-based and content-based layers respectively.

This demonstrates that the node layer activity can be thought of as being analogous to the 'vertical' aspect of node degree (where node degree would be 'horizontal').

Hypothesis 9.4 is rejected.

> ***Hypothesis 9.4: The multiplex node activity exhibits a power-law distribution with a negative exponent between*** $\mathbf{-2.5 \geq \gamma \geq -3.5}$**.**

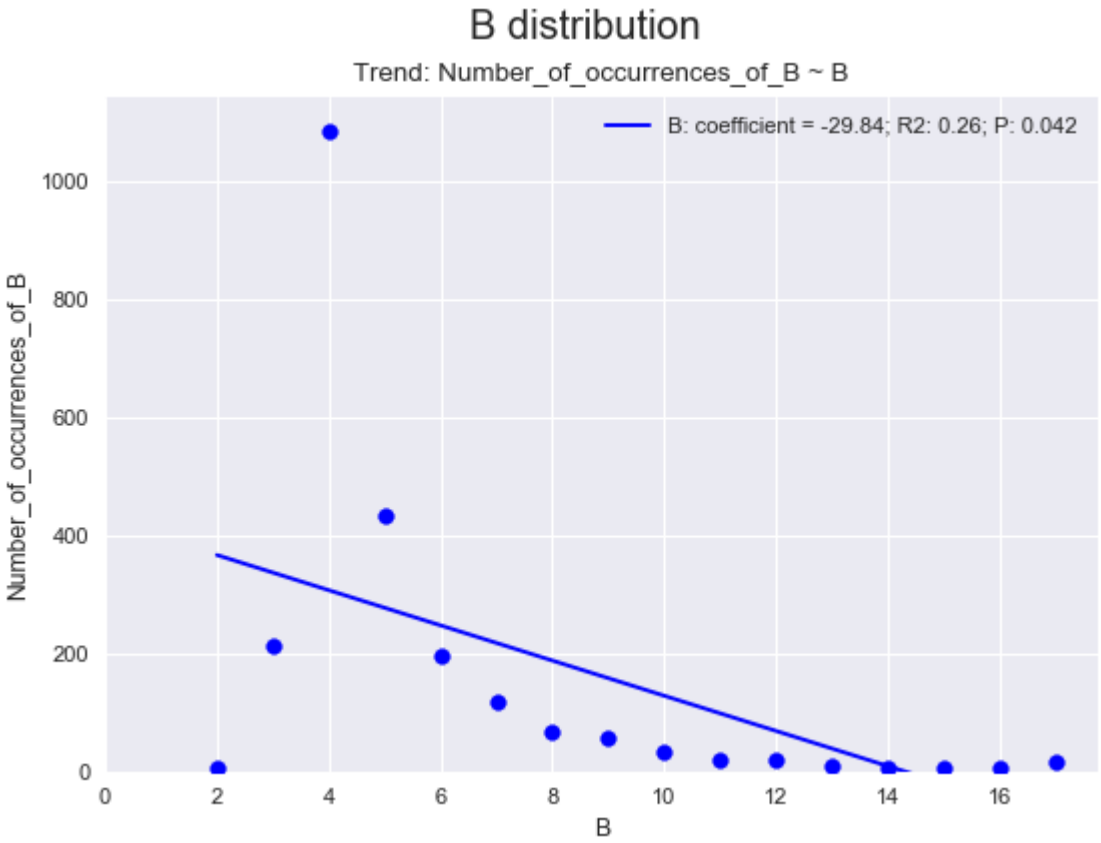

Figure 9.35. Node layer activity for the randomly assigned layers in a Barabási-Albert algorithm. It produces a wide fat-tailed Poisson distribution

9.5.5. Layer activity

The distribution of layer activity is given in Figure 9.36. Unlike the University of Bath multiplex co-authorship network, it does not provide a power-law relationship, and instead provides a Gaussian distribution, that is slightly skewed to the left as is suggested by the KDE plot.

This shape would likely produce a non-skewed Gaussian distribution if it were not for the interdisciplinary node entities providing a large rise in the number of node entities per layer (2,295 randomly distributed between 17 layers would suggest that mean would be 135 node entities per layer, instead of the ~520). This also explains why the distribution is skewed to the left as a high degree nodes would cause a large influx of new node entities.

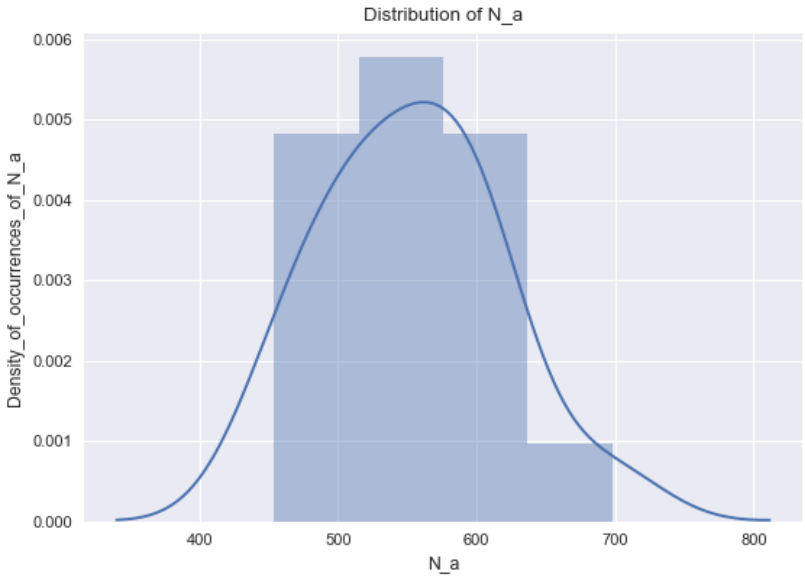

Figure 9.36 The distribution of layer activity for Model 2. It demonstrates a Gaussian distribution with a slight skew to the left.

9.5.6. Layer-pair closeness

The layer-pair closeness will be entirely reciprocal due to how it is that node entities are formed in this model. As such, it should form a Gaussian distribution. As can be seen in Figure 9.37, a Gaussian distribution suits the measure well. The University of Bath multiplex co-authorship network formed a clear power-law relationship and demonstrates how inaccurate this model is 'vertically'. This distribution implies that all layers have an equal probability to connect to all other layers.

Hypothesis 9.5 is rejected.

> ***Hypothesis 9.5: The multiplex layer-pair closeness exhibits a power-law distribution with a negative exponent between*** $-0.3 \geq \gamma \geq -1.0$.

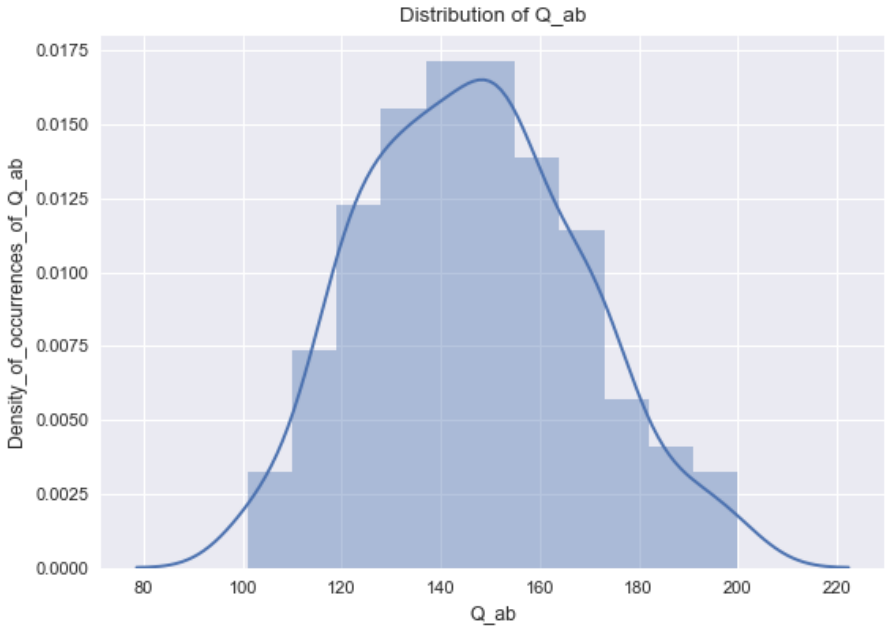

Figure 9.37. The layer-pair closeness distribution for Model 2 with a Gaussian distribution.

This is further confirmed by each of the 17 layers forming a normal distribution for the same of these layer-pair closeness values, as shown in Figure 9.38.

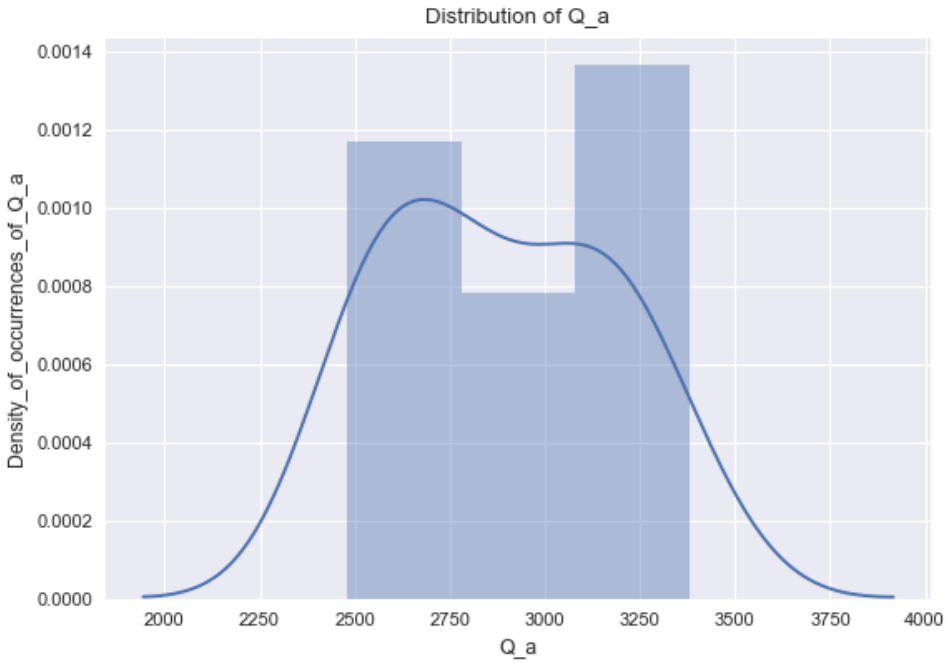

Figure 9.38. The distribution of layer-pair closeness activity for Model 2. It demonstrates a Gaussian distribution.

9.5.7. Discussion

This model provides a lens of features that are not considered realistic. The differences have been summarised in Table 9.7. The degree distribution provides reasonable approximations for the aggregate and all node entities distributions, as these represent the Barabási-Albert algorithm's contributions. Hypothesis 9.1 and 9.3(a) were corroborated, all other hypotheses were rejected. This suggests that the preferential attachment is a suitable mechanism. It is still subject to the criticisms of the Barabási-Albert algorithm that doesn't allow new nodes to become highly connected in comparison to older nodes. This could be overcome by introducing a fitness measure, aging measure, or link addition measures.

Disciplinary nodes are randomly selected, and by virtue of there being more poorly connected nodes, there is a bias towards choosing low degree nodes. High degree nodes therefore have a high probability of being present in most layers as interdisciplinary node entities. That such behaviour is not seen in real multiplex network structures shows that the barriers to IDR are being captured by the multiplex network framework.

The disciplinary degree vs. interdisciplinary degree provided as expected, a very unrealistic correlation, with the trend being non-linear, and greater than $k_{intra} = k_{inter}$ line. The non-linear trend provides an exemplar trend that favours lack of barriers.

The degree-correlations provided a very shallow, albeit a statistically significant negative power-law trend. The trend reversal in the aggregate network is not seen. The individual layer degree-correlations provide a higher magnitude exponent for its negative power-law trend. This is most likely due to the high degree nodes having many interdisciplinary node entities, thereby increasing the negative trend.

The most interesting results in comparison is how 'vertical' metrics differ from this model to the real-world results. In this respect, the total node activity produces very interesting results, as the simulated results suggest that it forms a fat-tailed Poisson distribution, with a sharp peak at four layers. Conversely, the University of Bath multiplex co-authorship network exhibits a strong negative power-law. As the degree distribution for a randomly connected Erdős-Renyi graph (Albert and Barabási 2002) also produces a Poisson distribution, where real networks produce a power-law relationship, there is a strong argument to be made that node activity in multiplex networks is analogous to the degree in traditional networks.

Equally, the layer activity, layer-pair closeness, and layer closeness centrality each produce Gaussian distributions across all layers in this random model. Conversely, the University of Bath multiplex co-authorship network exhibits negative power-law trends in the layer-pair closeness, and

a negative trend in the layer activity and layer closeness centrality (these are not statistically significant as they are based on only 17 layers).

Based on this information, it is proposed that the degree distribution, node activity, and layer-closeness be considered the main metrics for the network structure.

Table 9.7. Comparative values for the real-world department-based multiplex networks of the University of Bath and then Barabási-Albert algorithm model.

Measure		**Department-based multiplex networks**		**Barabási-Albert algorithm with randomly assigned core-disciplines**	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-1.87	$\log_{10} y \sim \log_{10} x$	-2.06
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.55	$\log_{10} y \sim \log_{10} x$	-2.25
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-1.20	$\log_{10} y \sim \log_{10} x$	-1.80
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-2.25	$\log_{10} y \sim \log_{10} x$	-4.30
Disciplinary-interdisciplinary boxplot		$y \sim x$	0.33	$y \sim f(x)$	>1
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	0.05*	$\log_{10} y \sim \log_{10} x$	-0.13
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.35	$\log_{10} y \sim \log_{10} x$	-0.44
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.28	$\log_{10} y \sim \log_{10} x$	-0.19*
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.36	$\log_{10} y \sim \log_{10} x$	-0.44*
B	**All nodes**	$\log_{10} y \sim \log_{10} x$	-3.32	Wide, fat-tailed Poisson distribution	Peak at 4 layers
N_a	**All nodes (only 4 points)**	$y \sim x$	-0.67*	Normal distribution	Mean: 535
Q_ab	**All nodes**	$\log_{10} y \sim \log_{10} x$	-0.65	Normal distribution	Mean: 142
Q_a	**All nodes (only 4 points)**	$y \sim x$	-12.02*	Normal distribution	Mean: 2780

*Not statistically significant.

**Significantly worse fit, lower R^2-value than its counterpart.

In comparison to Model 1 (the simultaneously grown Barabási-Model), there are no significant differences. The small differences between the two models are likely due to the random nature of randomly separating the nodes into layers in lieu of growing them on separate layers.

With regards to the analytical analysis, the random nature of the layer assignment makes it difficult to draw any further information with regards the distributions. This is because the layer assignment occurs at the end and not throughout the process.

However, as was pointed out in this model, the degree distribution, node activity, and layer-closeness are the most important quantities, and will therefore be represented in the analytical analyses in future models.

Table 9.8. A comparison of the Barabási-Albert simultaneous growth model to the University of Bath department-based multiplex co-authorship network.

Measure		**Barabási-Albert algorithm grown simultaneously on layers**		**Barabási-Albert algorithm grown simultaneously on layers**	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-2.76	$\log_{10} y \sim \log_{10} x$	-2.06
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.9; two-tail	$\log_{10} y \sim \log_{10} x$	-2.25
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-0.22	$\log_{10} y \sim \log_{10} x$	-0.13
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.23; two-tail	$\log_{10} y \sim \log_{10} x$	-0.44

Ultimately, the model exhibits completely unrealistic structures ‘vertically’. It is therefore strong evidence that barriers to IDR exist. Decision and policy makers could find it useful to see that both department-based and content-based divisions exhibit very significant barriers to collaboration (as concluded in section 9.3).

This evidence can be useful in highlighting what is actually happening instead of relying on people’s perception.

9.6. Model 3: Barabási-Albert with random edge assignment.

Drawing from the lessons of Models 1 and 2, it is possible to build up the growth model to address some of the weaknesses. Three major weaknesses were identified:

- Model 1 and 2 have no realistic mechanism to create interdisciplinary collaborations.
- The Barabási-Albert model's aggregate network degree distribution's exponent is too large in Model 1.
- No new links between existing nodes may be formed.

All three of these weaknesses may be addressed by introducing a mechanism where new links can be formed between node entities across all layers.

Therefore, an evolution component is required to complement the growth component. As with the first model, the Barabási-Albert algorithm is implemented on every layer simultaneously. The network is grown by layer, with a new node, i, introduced at every timestep on every layer with core-discipline D_i connecting to m_0 previously active node entities, j, on layer D_i. The node entities are chosen based on their degree given by $\Phi_{\mathrm{i}}^{\alpha}$.

$$\Phi_{\mathrm{i,t}}^{\alpha} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \tag{9.35}$$

This is useful as it is assumed every discipline grows on its own.

The second component requires a simple rule to ensure that interdisciplinary collaborations are possible. At every timestep, every node entity has a flat probability, $\Psi_{\mathrm{i}}^{\alpha}$, to link to m_1 previously active node entities. The other active node entities have equal probability, $\Theta_{\mathrm{i}}^{\alpha}$, to be assigned the other end of the links.

$$\Psi_{\mathrm{i,t}}^{\alpha} = C_0 \tag{9.36}$$

$$\Theta_{i,\mathrm{t}}^{\beta} = \frac{b_i^{\beta}}{\sum_{j=1}^{M} N_t^{\beta}} \cong \frac{1}{MN_t^{\beta}} \tag{9.37}$$

If the node entities are chosen for $\Psi_{\mathrm{i,t}}^{\alpha}$ and $\Theta_{\mathrm{j,t}}^{\beta}$ are i and j respectively, then a link between nodes (not node entities) i and j are added on layers α, D_i, and D_j respectively, but not on β. This is meant to mimic that of an individual who can have interdisciplinary collaborations, but they occur between the core-disciplines of the individuals involved.

9.6.1. Degree distribution by layer

This section establishes whether the Model 3's degree distribution matches the real-world network's degree distribution.

> ***Hypothesis 9.1: The degree distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.***

The aggregate degree distribution has an exponent of -2.44, as can be seen in Figure 9.39. This value is between the first two-models' values. It also appears to be a good fit throughout the distribution and does not appear to be skewed by the high degree, low occurrence nodes. This therefore corroborates Hypothesis 9.1(a).

> ***Hypothesis 9.1(a) - The aggregate degree distribution produces a power-law relationship with an exponent between -1.5 to -2.5***

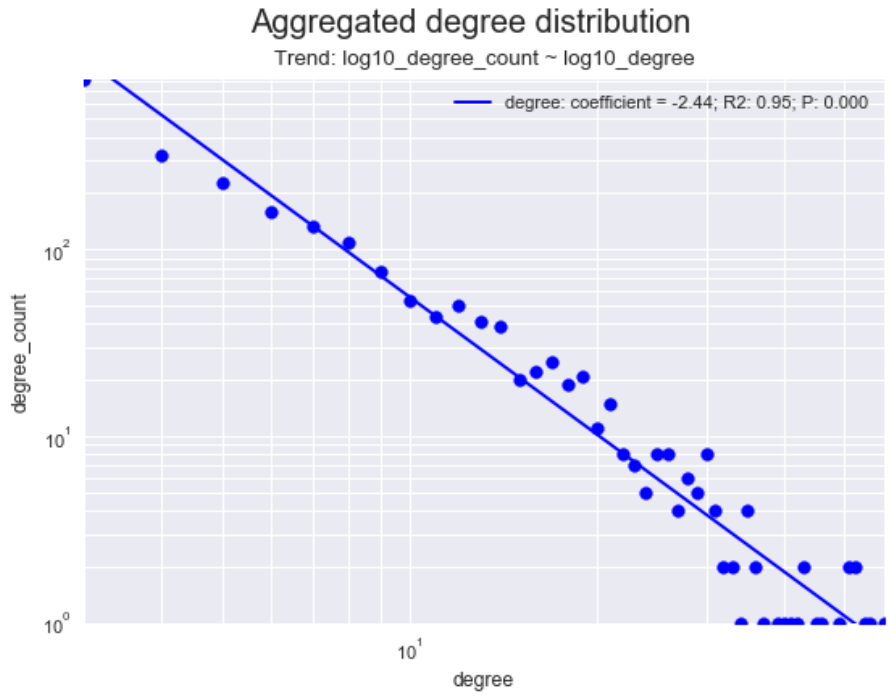

Figure 9.39. Aggregate degree distribution for Model 3.

The degree distribution by layer exhibits a strong power-law distribution on all, disciplinary, and interdisciplinary node entities as shown in Figure 9.40 to Figure 9.42. Therefore Hypothesis 9.1(b) is corroborated.

Hypothesis 9.1(b) - The degree distribution on every layer produces a power-law relationship using all node entities, disciplinary node entities only, and interdisciplinary node entities only.

These distributions are all strongly statistically significant but vary in magnitude. The disciplinary node entities show relatively few low degree nodes and many high degree nodes in comparison to its interdisciplinary counterpart. This is a very important result, as adding a random link in a traditional network, would be the equivalent of making the network a hybrid scale-free/Erdős-Renyi graph with both the scale-free and Poisson distribution being superimposed on each other. As there are ~3,800 disciplinary links and ~3,000 interdisciplinary links, there are more than enough links to make any Poisson distribution aspect prominent.

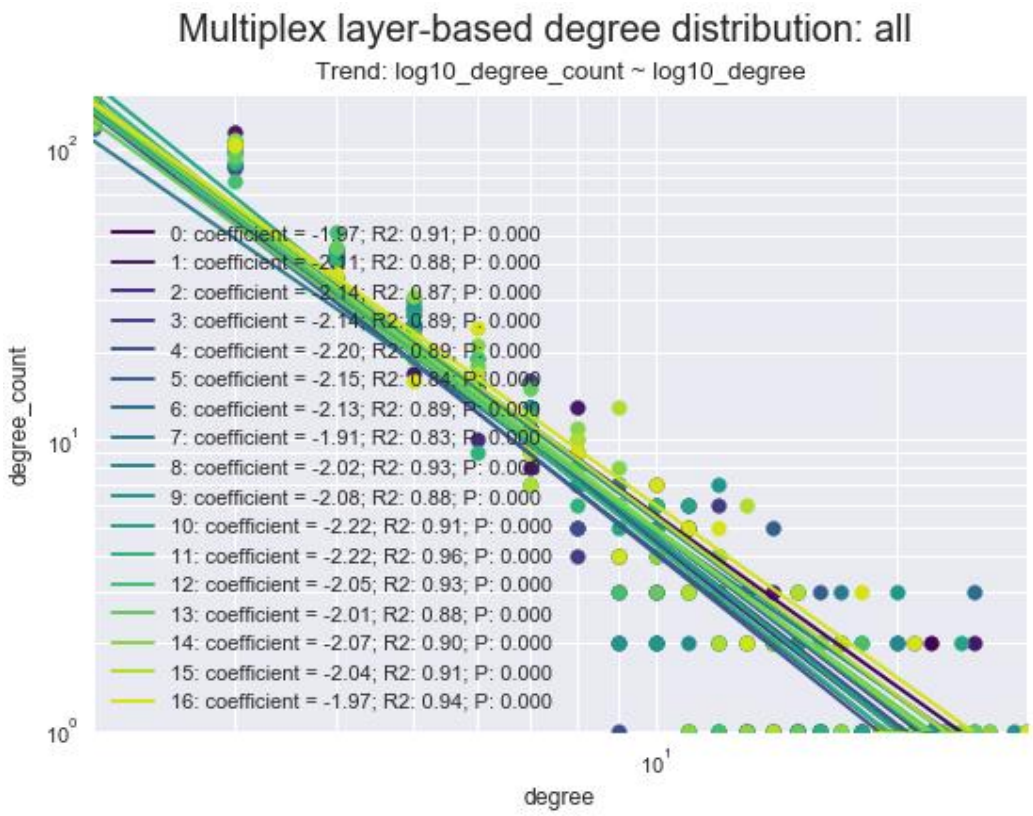

Figure 9.40. Layer degree distribution for all nodes for Model 3.

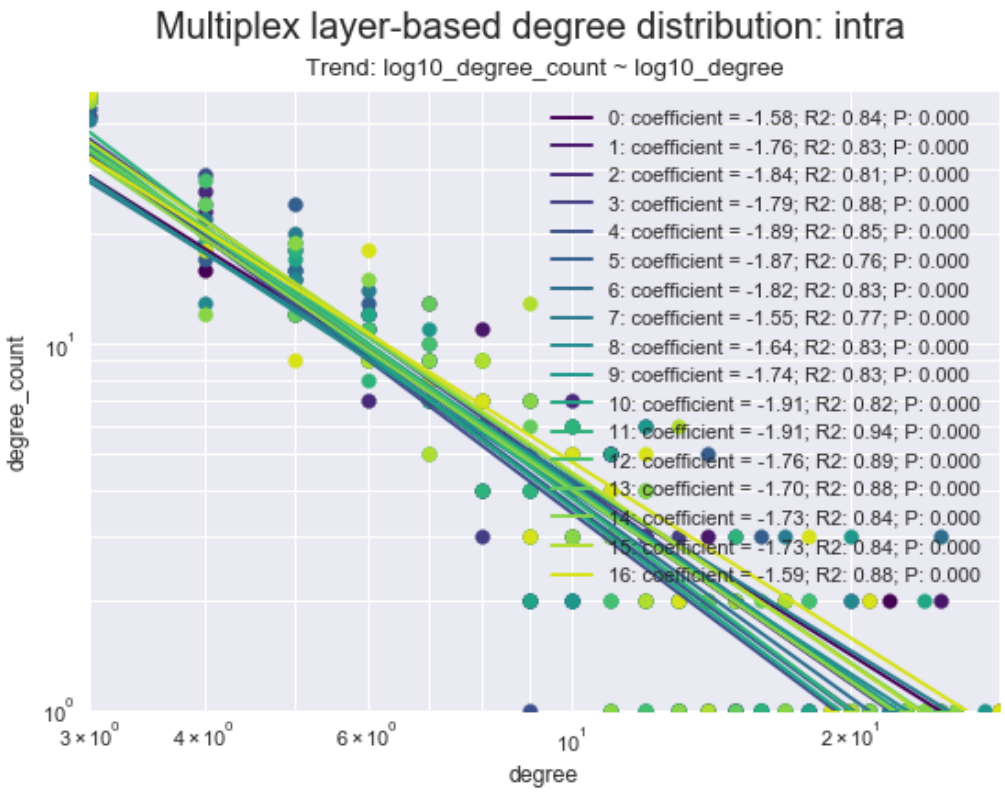

Figure 9.41.. Layer degree distribution for disciplinary nodes only for Model 3.

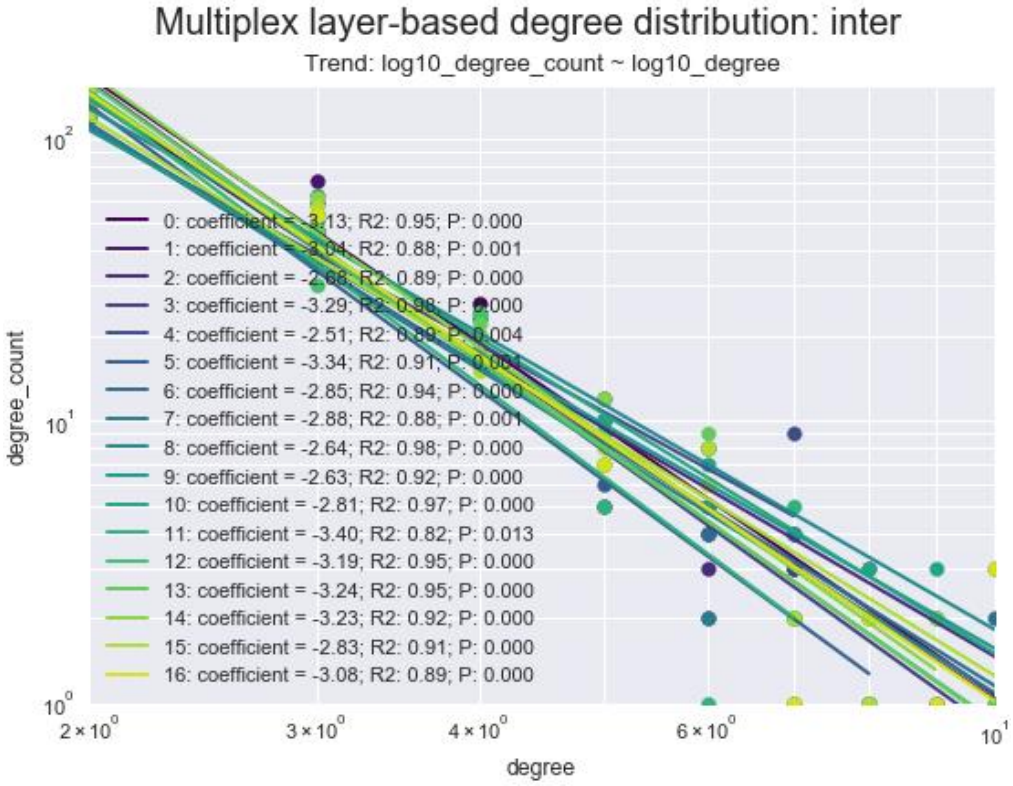

Figure 9.42. Layer degree distribution for interdisciplinary nodes only for Model 3.

The exponents form Gaussian distributions (Figure 9.43), with the interdisciplinary node entities providing the largest magnitude distribution exponents. The Gaussian distribution for all node entities and disciplinary node entities, are slightly skewed left, whereas the Gaussian distribution for the interdisciplinary node entities is skewed right. The peak densities occur at ~-2.14 for all node entities, ~-1.77 for disciplinary node entities, and ~-2.89 for interdisciplinary node entities. Therefore, Hypotheses 9.1(c)-(e) are corroborated. Hypothesis 9.1(f) is only corroborated for the interdisciplinary links, and rejected for the others. It is therefore only partially corroborated.

Hypothesis 9.1(c) - The degree distribution on every layer, using every node entity, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than the aggregate exponent.

Hypothesis 9.1(d) - The degree distribution on every layer, using disciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than all the node entities' peak exponent.

Hypothesis 9.1(e) - The degree distribution on every layer, using interdisciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent above the aggregate exponent.

Hypothesis 9.1(f) - The degree distributions' exponents are distributed as Gaussians that are skewed to the right as estimated by the KDE.

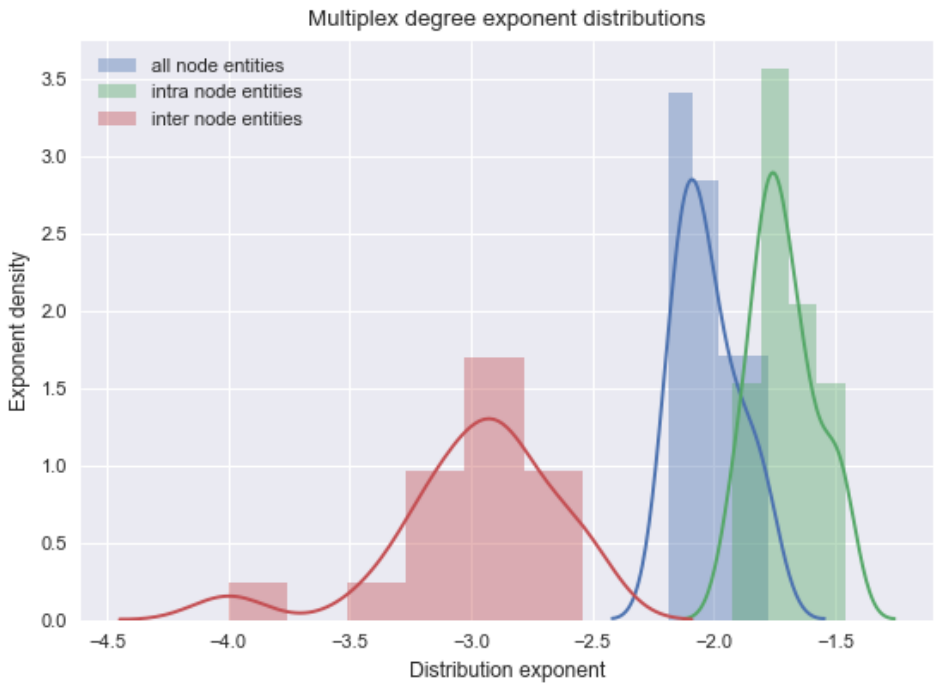

Figure 9.43. Layer power-law exponent distributions for Model 3. Disciplinary nodes are skewed left, whilst interdisciplinary nodes are skewed right.

The differences between disciplinary and interdisciplinary node entities occur since there are three mechanics at play: preferential attachment happening within the layer according to $\Phi^{\alpha}_{i,t}$, random connections occurring for node entities with probability $\Psi^{\alpha}_{i,t}$, and a probability to connect to all nodes' core-discipline node entities $\Theta^{\alpha}_{i,t}$.

Therefore, disciplinary nodes are guaranteed a starting degree of $k^{\alpha}_{i}(t_i) = m_0$, whereas interdisciplinary nodes will only ever start with $k^{\alpha}_{i}(t_i) = 1$. This may not seem like much, but due to the rich-get-richer phenomenon, it is likely to have a significant effect, as disciplinary nodes are m_0 times as likely from the outset to get new connection from $\Phi^{\alpha}_{i,t}$. This mechanism causes disciplinary nodes to have a higher proportion of higher degree nodes, reducing the exponent. This can also be seen in that there are far more low degree nodes in interdisciplinary degree distributions compared to the disciplinary degree distributions, and vice-a-versa for high degree nodes. Equally, most of the nodes are disciplinary, and are therefore more likely to receive $\Theta^{\alpha}_{i,t}$ links, which exacerbates this phenomenon.

Whilst the skewness was not perfectly matched for the disciplinary node entities, Hypothesis 9.1 is almost entirely corroborated.

A reasonable power-law structure is simulated when a major component of the simulation is to randomly allocate links. This is an important finding as the scale-free property of real networks

have shown to be distinctly different from randomly connected networks (which exhibit a Poisson distribution).

Upon further investigation, it was discovered that it was due to the existence of node entities that enables this. The more active node entities a node has, the more likely they are to be selected, which in itself is a rich-get-richer mechanism.

The implication (or systems theory) of this is that node entities play an important role in collaboration. Specifically, a person's interdisciplinarity (i.e. the node activity) is an important factor in establishing new IDR collaborations.

For instance, if person A is active in 5 disciplines, person B is active in 10 disciplines, and person C is active in 15 disciplines, then the probability of person D conducting IDR with C is greater than with B, which is greater than with A (assuming they're all from different disciplines). This probability drives underlying mechanics in the network that ensures that causes power-law degree distributions to be seen in all layers and in the aggregate network such that it mimics real network structures.

Furthermore, a node with an established presence in another discipline will start reaping the rewards of traditional growth and development of a field. That is to say, they have established themselves as a researcher who is relevant in that discipline. The more prominent they are, the more likely they are to draw new collaborators in that discipline.

However, this only holds true for a specific discipline. If the researcher wants to establish themselves in another discipline, prominence in any or multiple other disciplines has no evidence of being beneficial. This could perhaps be indicative a person's research interests – i.e. if they have established themselves as interdisciplinary researchers, then they are more likely to conduct IDR in comparison to someone who has researched a lot in two different disciplines.

Therefore, to enter a new discipline, interdisciplinarity is important. To sustain collaborations in that new discipline, gaining prominence in that discipline is vital.

This provides decision and policy-makers with a useful model. If they seek to enable IDR, they should seek interdisciplinary individuals, if they seek to sustain IDR between two different disciplines, they should find the individual who has interdisciplinary prominence.

9.6.2. Disciplinary vs interdisciplinary degree regression.

Comparing the nodes disciplinary degree to the sum of the interdisciplinary degree, it can clearly be seen in Figure 9.44 that the probability chosen for $\Psi_{i,t}^{\alpha}$ and $\Theta_{i,t}^{\alpha}$ have not skewed the results and are therefore proportional. If the probability were higher, the mean line would be above the $k_{intra} =$

k_{inter} line. What is interesting however, is the occurrence of low k_{intra} nodes with higher k_{inter}, despite the favouritism that is given to interdisciplinary node entities. This is because when these nodes establish early footholds in given interdisciplinary layers, they benefit from the rich-get-richer phenomenon.

Hypothesis 9.2 is corroborated.

Hypothesis 9.2: The disciplinary node entities degrees are larger than the median of the sum of their counterpart interdisciplinary node entities' degrees.

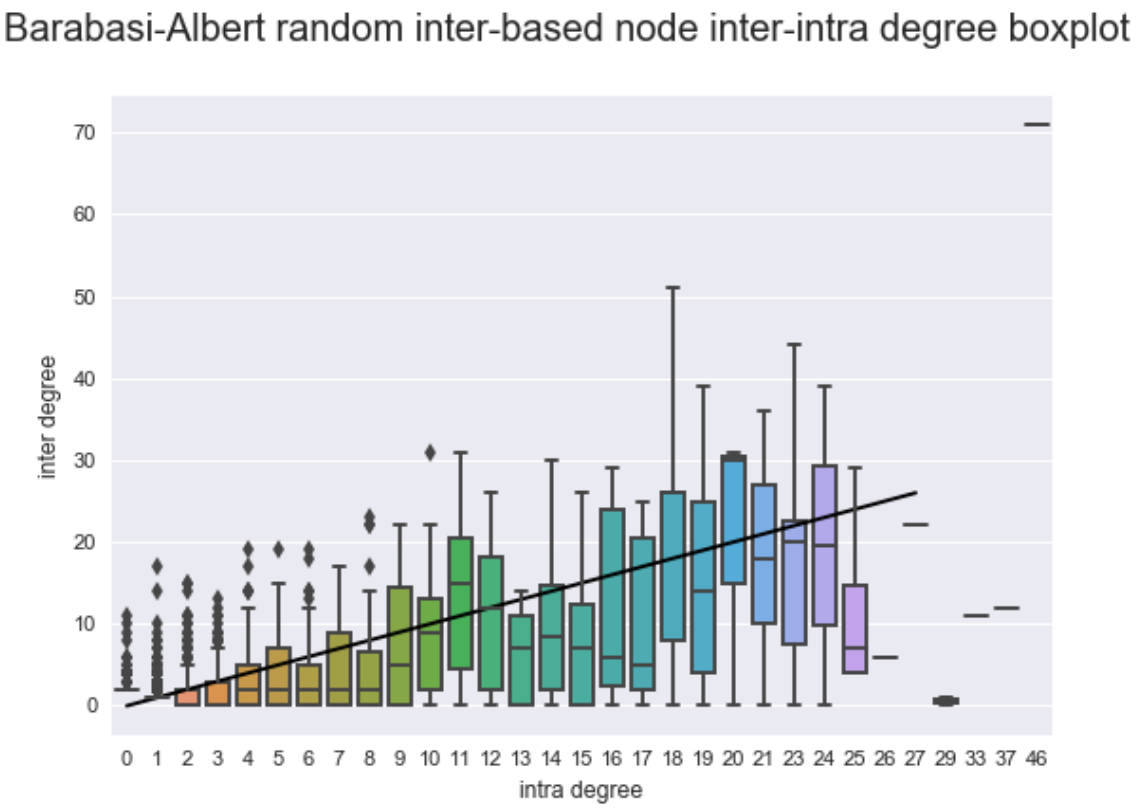

Figure 9.44. Disciplinary node entities degree vs. the interdisciplinary node entities' sum of degrees for Model 3.

9.6.3. Degree-correlations

This section establishes whether the degree-correlations in this model match the degree-correlations of the real-world network.

Hypothesis 9.3: The degree-correlation distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.

The aggregate degree-correlation in this model produces a statistically significant negative trend as can be seen in Figure 9.45. This disagrees with the results from the University of Bath multiplex co-authorship network.

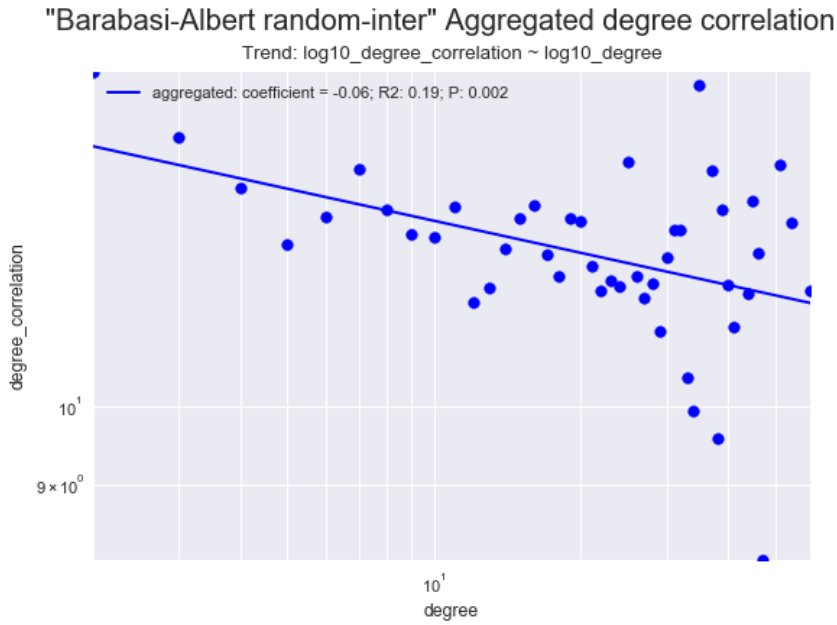

Figure 9.45. The aggregate degree-correlation for Model 3 is statistically significant with a negative trend.

The layer degree distribution for all node entities, disciplinary node entities, and interdisciplinary node entities are shown in Figure 9.46 to Figure 9.48 respectively. It can clearly be seen that a power-law distribution suits each but exhibit a lot of noise. Furthermore, many of the distributions are not statistically significant. All the statistically significant layers do exhibit a power-law relationship with a negative exponent. Hypothesis 9.3(a) is partially corroborated.

Hypothesis 9.3(a) - Layers exhibit degree-correlation distributions with a power-law relationship with a negative exponent.

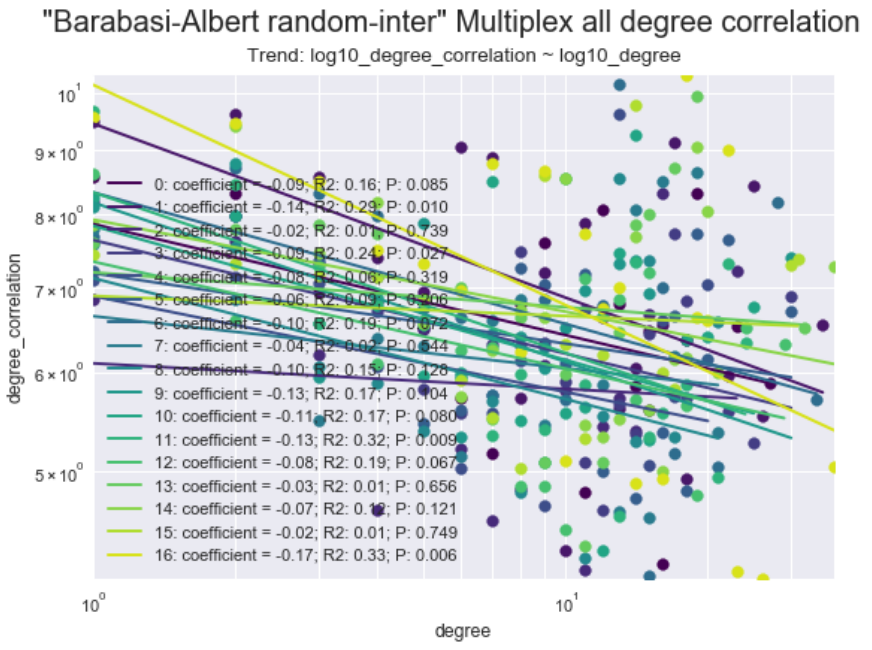

Figure 9.46. The degree-correlation for Model 3 for all nodes with random edge assignments.

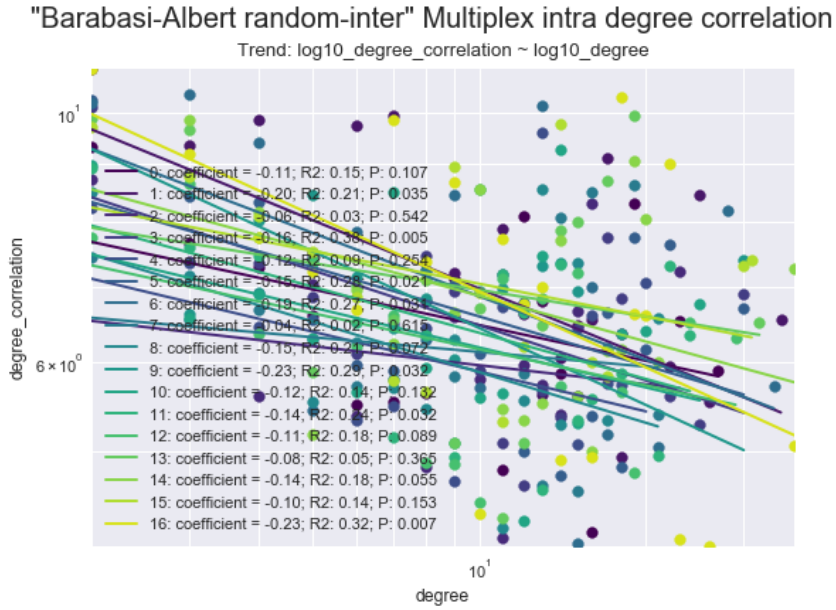

Figure 9.47. The degree-correlation for Model 3 for disciplinary nodes only with random edge assignments.

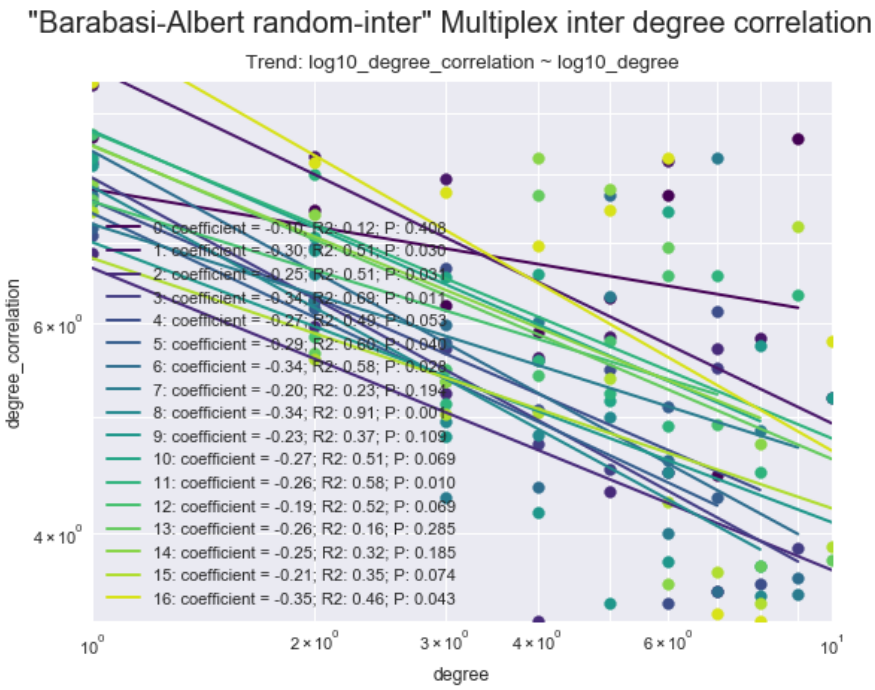

Figure 9.48. The degree-correlation for Model 3 for interdisciplinary nodes only with random edge assignments.

The distribution peaks are all above 0.3, thereby corroborating Hypothesis 9.3(b).

Hypothesis 9.3(b) - Degree-correlation distribution exponents exhibit Gaussian distributions as estimated by the KDE skewed right with the peak density occurring at a value of $\gamma > 0.3$.

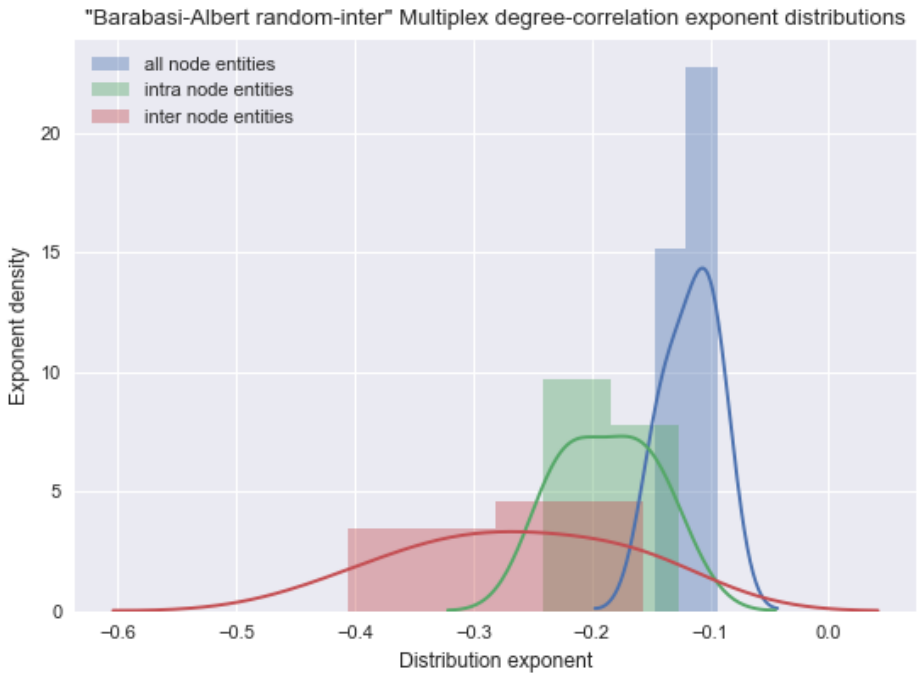

Figure 9.49. The degree-correlation exponent distributions of Model 3's network. This matches well with the real-world network results.

Assuming the reversal in trend from aggregate to layer is true; this was not reproduced, and this model is unable to explain why this may occur. Despite this, Hypothesis 9.3 is corroborated.

9.6.4. Node activity

At a glance, the node activity would be expected to form a Poisson distribution as the nodes and layers in which they become active are randomly chosen. However, the probability is not chosen based on the nodes, but rather the node entities. This means that probabilities of forming a new link and being chosen to receive the new link are given in the following expressions respectively.

$$\Psi_i = \sum_{\alpha=1}^{M} \Psi_i^{\alpha} \tag{9.38}$$

$$\Theta_i = \sum_{\alpha=1}^{M} \Theta_i^{\alpha} \tag{9.39}$$

As both are summed by the number of layers, the more node entities that exist, the higher the probability of getting new links will be. Therefore, early active nodes have the same advantage as they do with preferential attachment, $\Phi_{i,t}^{\alpha}$.

Therefore, the node activity distribution can form a power-law relationship. The simulation results in Figure 9.50 show that a power-law distribution with an exponent of -2.73 was created. Therefore, Hypothesis 9.4 is corroborated.

Hypothesis 9.4: The multiplex node activity exhibits a power-law distribution with a negative exponent between $-2.5 \geq \gamma \geq -3.5$.

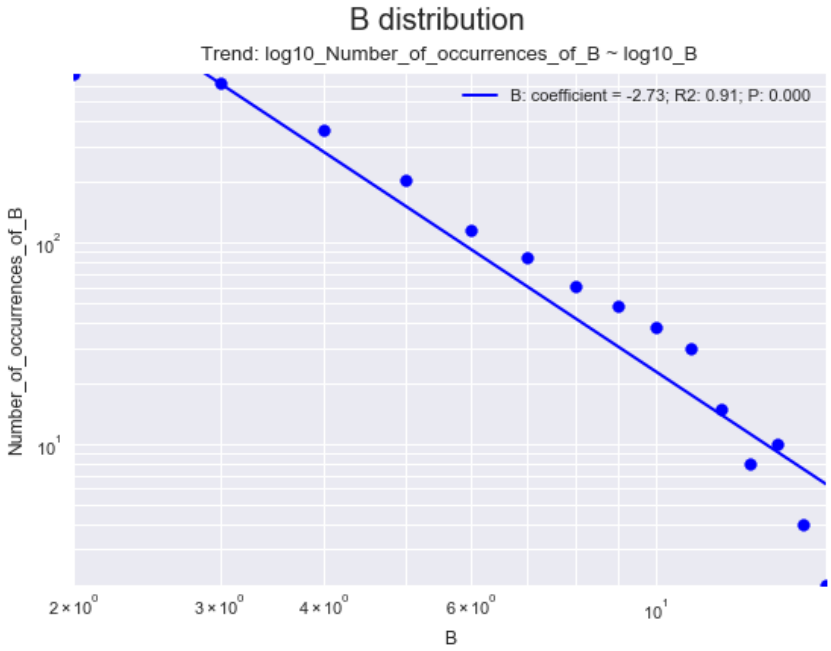

Figure 9.50. Node activity distribution for Model 3. This is statistically significant with a strong negative correlation.

The most important aspect of this model is that greater node activity essentially creates a preferential attachment mechanism that ensure random connections creating a power-law degree distribution, but also a power-law node activity. This implies that the more interdisciplinary a researcher is, the more likely it is that they will become more interdisciplinary.

This further supports the fact that a preferential attachment is occurring due to node entities. Again, this implies that on a nodular basis, this model performs quite well. It therefore highlights the importance of node activity to establish new IDR, and the prominence of interdisciplinary research in a specific discipline to sustain such IDR.

9.6.5. Layer activity

Despite the power-law distribution exhibited in the node activity, the same phenomenon should occur for layers. A layer whose core nodes have many node entities is more likely to be selected

when new links are added. However, as there are a fixed number of layers, which all exist at the same time, this advantage cannot be taken fully advantage of.

As such, the resulting distribution of active nodes per layer yields a normal distribution as shown in Figure 9.51, with the peak density occurring at ~410 nodes. This is a poor fit for the University of Bath multiplex co-authorship network.

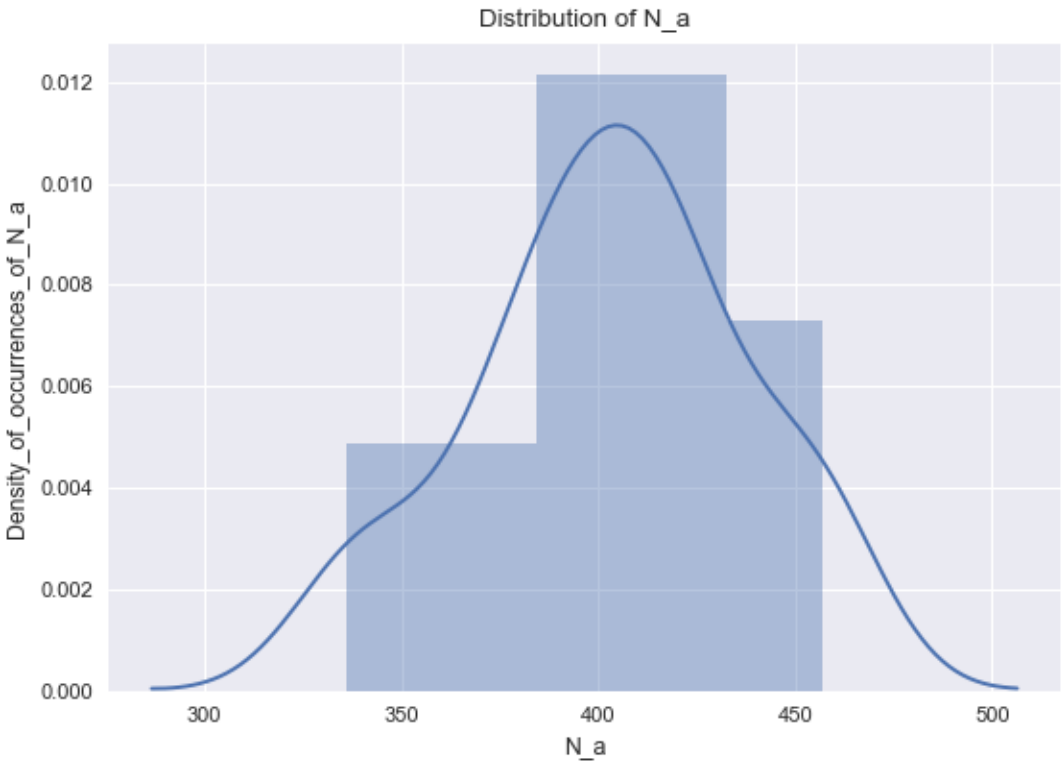

Figure 9.51. Distribution of the number of active nodes per layer for Model 3.

9.6.6. Layer-pair closeness

The layer-pair closeness distribution matches a Gaussian distribution relatively well with the mean occurring at roughly 95 nodes being active in two layers as seen in Figure 9.52. The spread is quite large, and the mode and density peak occurs at ~105 nodes. This is a poor fit for the University of Bath multiplex co-authorship network. Furthermore, there is no specific layer that breaks this mould, as can be seen in Figure 9.53. Hypothesis 9.5 is therefore rejected.

> ***Hypothesis 9.5: The multiplex layer-pair closeness exhibits a power-law distribution with a negative exponent between*** $-0.3 \geq \gamma \geq -1.0$.

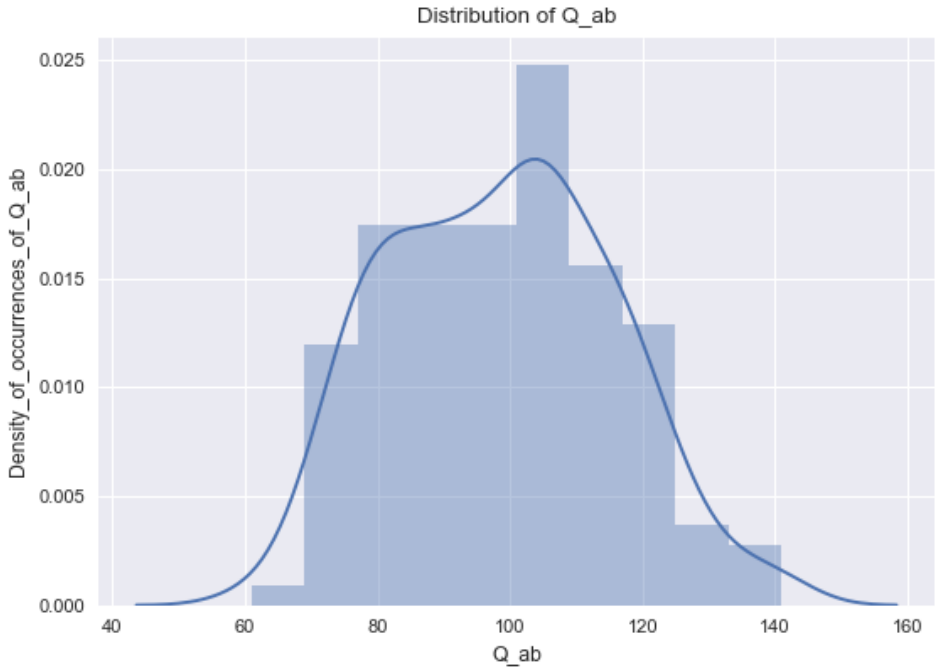

Figure 9.52. Distribution of the number of co-active nodes per layer pair for Model 3.

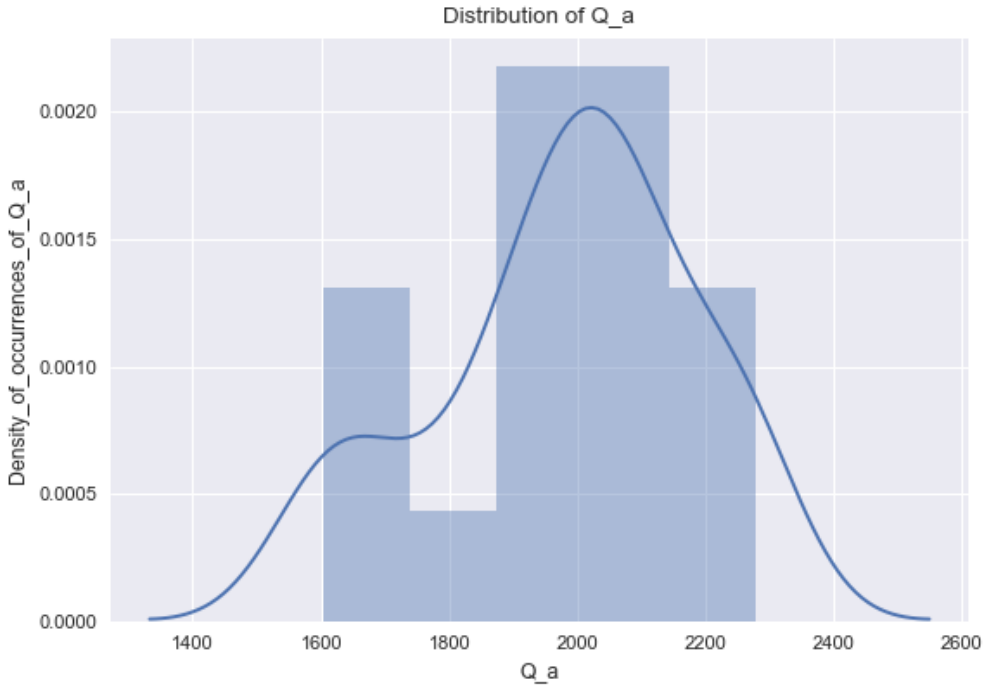

Figure 9.53. Distribution of the sum of number of co-active nodes per layer pair for every layer for Model 3.

9.6.7. Analytical analysis

As was outlined in the discussion of Model 2, three different variables are of interest: the degree, node activity, and layer-closeness.

The distributions of these provide vital aspects of the multiplex structure. The distributions should all follow a power-law relationship.

The analytical solutions for the distributions are very difficult to express as they rely on the proportion of interdisciplinary nodes in each layer. This is a very difficult quantity to develop an expression for. An approximate solution for the degree distribution can be found in Appendix.

A far simpler and potentially more useful analysis is to develop an expression for the rate of change of these variables, as can give an indication as whether heterogeneous distributions can occur. More importantly however, the rates of change expressions provide predictive capability (assuming linearization).

The following expressions recap the mechanics adopted for this model.

$$\Phi_{\mathrm{i,t}}^{\alpha} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \tag{9.40}$$

$$\Psi_{\mathrm{i,t}}^{\alpha} = C_0 \tag{9.41}$$

$$\Theta_{\mathrm{j,t}}^{\beta} = \frac{1}{\sum_{j=1}^{M} N_t^{\beta}} \cong \frac{1}{MN_t^{\beta}} \tag{9.42}$$

This section develops expressions for $\frac{dk_i^{\alpha=D_i}}{dt}$ and $\frac{dk_i^{\alpha\neq D_i}}{dt}$. There are three mechanisms by which a node can increase its degree:

- By connecting to a new node.
- By creating a new link to an old node.
- By receiving a new link from an old node.

The difference between the disciplinary and interdisciplinary node entities' degrees is that disciplinary degrees benefit from the presence of all node entities.

This results in the following expressions.

$$\frac{dk_i^{\alpha=D_i}}{dt} = m_0\Phi_{\mathrm{i,t}}^{\alpha} + B_i m_1 \Psi_{\mathrm{i,t}}^{\alpha} + m_1 \sum_{\beta}^{M} \left(\sum_{j=1}^{N_t^{\beta}} \left(\Psi_{j,\mathrm{t}}^{\beta} \right) \right) \cdot \sum_{\beta}^{M} \left(B_i \Theta_{i,\mathrm{t}}^{\beta} \right) \tag{9.43}$$

$$\frac{k_i^{\alpha \neq D_i}}{dt} = m_0 \Phi_{\mathrm{i,t}}^{\alpha} + m_1 \Psi_{\mathrm{i,t}}^{\alpha} + m_1 \sum_{\beta}^{M} \left(\sum_{j=1}^{N_t^{\beta}} \left(\Psi_{j,\mathrm{t}}^{\beta} \right) \right) \cdot \sum_{\beta}^{M} \left(\Theta_{i,\mathrm{t}}^{\beta} \right) \tag{9.44}$$

Solving for equation 9.4.

$$\frac{dk_i^{\alpha = D_i}}{dt} = m_0 \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + B_i m_1 C_0 + m_1 \sum_{\beta}^{M} \left(\sum_{j=1}^{N_t^{\beta}} (C_0) \right) \cdot \sum_{\beta}^{M} \left(\frac{B_i}{M N_t^{\beta}} \right) \tag{9.45}$$

$$\frac{dk_i^{\alpha = D_i}}{dt} = m_0 \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + B_i m_1 C_0 + m_1 C_0 M \langle N_t^{\beta} \rangle \cdot \frac{B_i}{\langle N_t^{\beta} \rangle} \tag{9.46}$$

$$\frac{dk_i^{\alpha = D_i}}{dt} = m_0 \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + B_i m_1 C_0 + B_i m_1 C_0 M \tag{9.47}$$

$$\frac{dk_i^{\alpha = D_i}}{dt} = m_0 \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + m_1 C_0 (1 + M) B_i \tag{9.48}$$

Solving for equation 9.5

$$\frac{dk_i^{\alpha \neq D_i}}{dt} = m_0 \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + m_1 C_0 (1 + M) \tag{9.49}$$

This analytical expression shows why the random links do not create a Poisson distribution for disciplinary node entities. The reason is that B_i is scale-free, and so the seemingly random connections are multiplied by a scale-free property.

Interdisciplinary node entities do not gain this benefit. Interdisciplinary node entities are more susceptible to getting a Poisson distribution as there is nothing mitigating the scale-free component being overlapped by the random component.

The rate of change of B_i is given in the following expression (assuming that $B_i \ll M$).

$$\frac{dB_i}{dt} = B_i \left(m_1 \Psi_{\mathrm{i,t}}^{\alpha} \sum_{\beta \neq \alpha}^{M} \Theta_{j,\mathrm{t}}^{\alpha\beta} + m_1 \sum_{\beta \neq \alpha}^{M} \sum_{j \neq i}^{N^{\beta}} \Theta_{i,\mathrm{t}}^{\beta\alpha} \Psi_{\mathrm{j,t}}^{\beta} \right) \tag{9.50}$$

$$\frac{B_i}{dt} \cong m_1 C_0 {B_i}^2 \left(\frac{1}{\langle N_t^\beta \rangle} + 1 \right) \tag{9.51}$$

The node activity is guaranteed to increase due to the $\left(\frac{1}{\langle N_t^\beta \rangle} + 1 \right)$ term. However, the scale-free property is guaranteed by virtue of the ${B_i}^2$ term.

The layer pair closeness can be approximated as follows.

$$\boldsymbol{\frac{dQ_{\alpha\beta}}{dt} = m_1 \sum_{j \neq i}^{N^\alpha} \Psi_{\mathrm{i,t}}^{\alpha} \sum_{j \neq i}^{N^\beta} \Theta_{\mathrm{j,t}}^{\alpha\beta} + m_1 \sum_{j \neq i}^{N^\alpha} \Theta_{i,\mathrm{t}}^{\beta\alpha} \sum_{j \neq i}^{N^\beta} \Psi_{\mathrm{j,t}}^{\beta}} \tag{9.52}$$

$$\frac{dQ_{\alpha\beta}}{dt} = m_1 \frac{m_1 C_0 N^\alpha N^\beta}{M} + \frac{m_1 C_0 N^\alpha N^\beta}{M} \tag{9.53}$$

$$\frac{dQ_{\alpha\beta}}{dt} = 2 \frac{{m_1}^2 C_0 N^\alpha N^\beta}{M} \tag{9.54}$$

$\frac{dQ_{\alpha\beta}}{dt}$ is therefore dependent on $N^\alpha N^\beta$. It is difficult to determine whether this should provide a power-law distribution without developing expressions for $N^\alpha N^\beta$.

9.6.8. Discussion

Despite its simplicity, this model provided results that matched the trends of the University of Bath multiplex co-authorship network relatively well. The two are compared in Table 9.9. The disciplinary degree vs. interdisciplinary degree was below the $k_{intra} = k_{inter}$ line, suggesting that the model was reasonably tuned.

The degree distribution provided entirely power-law distributions, despite the random nature of the link addition. This is a very significant finding. Random connections between nodes in traditional networks always results in a strong Poisson distribution type of connections. This is not a case of it simply being lost within the Barabási-Albert model, as the number of random links added were significant. This model created ~3,800 disciplinary links and ~3,000 interdisciplinary links. If a Poisson distribution were prominent, it would be seen in the results (Note: as the probability of $\Psi_{\mathrm{i,t}}^{\alpha} \to 1$, a Poisson distribution is guaranteed).

This is because the probability is tied to the activity and is not truly completely random, but rather subject to following expression.

$$\Psi_{\mathrm{i}} = \sum_{\alpha=1}^{M} \Psi_{\mathrm{i}}^{\alpha} = B_i C_0 \tag{9.55}$$

$$\Theta_{\mathrm{i}} = \sum_{\alpha=1}^{M} \Theta_{\mathrm{i}}^{\alpha} \cong \frac{1}{MN_t^{\beta}} \tag{9.56}$$

It is this same mechanism alongside the growth of the network that makes B_i scale-free, and therefore the degree distribution scale-free. Therefore, the multiplex aspect of node entities mimics the preferential attachment, contributing towards the scale-free properties of the multiplex network.

Contribution to knowledge

Multiplex growth models that grow based on node entities' degrees are resilient to random connections. Despite a large number of the links added being entirely randomly, there is little evidence of the expected Poisson distribution occurring in degree distributions in any level of aggregation.

This is entirely due to the existence and the inclusion of the node entities in the model. The node entities drive a preferential attachment mechanism by virtue of random connections being more likely to be attached to nodes with many active node entities.

This then recreates individual node properties very well using a very simple model. Most notably, it produces a power-law distribution for both the degree and the node activity.

The implication of this is that node entities play a central role in recreating accurate multiplex structures. The implication of the model itself, by virtue of every node entity attracting links on its own is that individuals' presence in other disciplines can be treated as semi-independent people. That is to say, it matters little what a person's status in another discipline is, and is very dependent on what their status in the discipline of interest is (e.g. if person A is highly regarded in Mechanical Engineering, they are not guaranteed to grow in Management. If person B is average in Mechanical Engineering and in Management, they are likely to outperform person A in Management, and underperform person A in Mechanical Engineering). This provides a very tribal view of IDR, where breaking the barriers are very difficult, but once you do, you stand to gain from the same principles of the rich-get-richer as everyone else in that discipline.

This represents a significant contribution to knowledge as this behaviour has not been previously identified.

However, there are weaknesses to the model, the foremost being the unrealistic distribution of the layer-pair closeness. This is a very important multiplex structural measure that defines how it is that nodes and layers are connected. This has not been mimicked at all in Model 3.

This also does not represent the distribution of layers well either, but this could be defined by the University (i.e. there are more people available in Physics than in Management due to University policy).

Furthermore, the trend reversal in the degree-correlation has not been achieved.

The model was not intended to be a good representation at the outset, but the strong scale-free dynamics driven by the link addition and multiplex node entities synergy provide a deep insight into multiplex structures. As a result, the degree distribution, disciplinary vs. interdisciplinary degree comparison, degree-correlation, and node activity hypotheses were all mostly corroborated. It was only the layer-pair closeness hypothesis that was fully rejected.

It has to be pointed out however, that only the node activity measure was a good match out of the 'vertical' measures. Furthermore, the exponents were too large in magnitude for the degree distributions.

However, this model provides no improvement on modelling the layer-pair closeness, representing the final frontier of being able to recreate the real-world multiplex structure.

Table 9.9. Comparative values for the real-world department-based multiplex networks of the University of Bath and Model 3.

Measure		Department-based multiplex networks		Barabási-Albert algorithm with randomly assigned core-disciplines	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-1.87	$\log_{10} y \sim \log_{10} x$	-2.44
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.55	$\log_{10} y \sim \log_{10} x$	-2.15
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-1.20	$\log_{10} y \sim \log_{10} x$	-1.77
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-2.25	$\log_{10} y \sim \log_{10} x$	-2.91
Disciplinary-interdisciplinary boxplot		$y \sim x$	0.33	$y \sim f(x)$	<1
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	0.05*	$\log_{10} y \sim \log_{10} x$	-0.06
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.35	$\log_{10} y \sim \log_{10} x$	-0.11*
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.28	$\log_{10} y \sim \log_{10} x$	-0.19*
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.36	$\log_{10} y \sim \log_{10} x$	-0.28*
B	**All nodes**	$\log_{10} y \sim \log_{10} x$	-3.32	$\log_{10} y \sim \log_{10} x$n	-2.73
N_a	**All nodes (only 4 points)**	$y \sim x$	-0.67*	Normal distribution	410
Q_ab	**All nodes**	$\log_{10} y \sim \log_{10} x$	-0.65	Normal distribution	105
Q_a	**All nodes (only 4 points)**	$y \sim x$	-12.02*	Normal distribution	2020**

*Not statistically significant.

**Poor fit.

Some further mechanics can be noted.

It has been reported that the double preferential attachment reduces the degree distributions' exponents (Ghoshal, Chi et al. 2013). As can be seen in Table 9.10, the degree distributions' exponents are smaller in Model 3 than in Model 2. This provides further evidence that the node entities provide preferential attachment that make the model resilient to random connections.

Contribution to knowledge

In a model with a large component of randomly connected links, the degree distribution is expected to exhibit a Poisson distribution. However, in multiplex networks, this effect is dampened out. This is because a node will have multiple node entities (provided they can only connect to active node entities). The node entities therefore increase the probability of a node being connected to.

This is a preferential attachment mechanism centred on the node activity.

The importance of the node activity and its natural tendency to cause preferential attachment to occur solidify node activity as being akin to a 'vertical' node degree.

Table 9.10 Comparative values for Model 3 to Model 3 with double preferential attachment.

Measure		Barabási-Albert algorithm with randomly assigned links to node entities		Barabási-Albert algorithm with links assigned to nodes based on node degree	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-2.44	$\log_{10} y \sim \log_{10} x$	-2.14
	All node entities	$\log_{10} y \sim \log_{10} x$	-2.15	$\log_{10} y \sim \log_{10} x$	-1.73
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-1.77	$\log_{10} y \sim \log_{10} x$	-1.50
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-2.91	$\log_{10} y \sim \log_{10} x$	-3.83
Disciplinary-interdisciplinary boxplot		$y \sim x$	<1	$y \sim x$	<1
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-0.06	$\log_{10} y \sim \log_{10} x$	-0.13
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.11*	$\log_{10} y \sim \log_{10} x$	-0.29
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.19*	$\log_{10} y \sim \log_{10} x$	-0.26
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.28*	$\log_{10} y \sim \log_{10} x$	-0.53*
B	**All nodes**	$\log_{10} y \sim \log_{10} x$n	-2.73	$\log_{10} y \sim \log_{10} x$n	-3.52
N_a	**All nodes (only 4 points)**	Normal distribution	410	Normal distribution	272
Q_ab	**All nodes**	Normal distribution	105	Normal distribution	37
Q_a	**All nodes (only 4 points)**	Normal distribution	2020**	Normal distribution	847

*Not statistically significant.

**Poor fit.

It is worth noting that attempts at mimicking the layer-pair closeness included the measure itself, and yielded no further improvements to the model.

$$\Psi_{\mathrm{i}} = \left(1 + Q_{\alpha\beta} k_i\right) C_0 \frac{k_i}{\sum_{j=1}^{N_t} k_j} \tag{9.57}$$

$$\Theta_{\mathrm{i}} = \frac{k_i}{\sum_{j=1}^{N_t} k_j} \tag{9.58}$$

9.7. Model 4: Barabási-Albert model with links addition based on layer closeness centrality and single preferential attachment

Model 1 was able to capture the 'horizontal' structural aspects of real multiplex networks. Model 3 was able to capture individual nodes 'vertical' aspects (i.e. node activity). However, no model was able to capture the overall 'vertical' structure of real multiplex networks, the most important measure of which is the layer closeness.

Therefore, one further rule was implemented, the layers themselves are analogous to nodes, with every common link between two layers counting as an increase in weight. This would make $Q_{\alpha\beta}$ the layer link weight, and $Q_\alpha = \sum_{\beta=1}^{M} Q_{\alpha\beta}$ the strength of the layer.

A rule can therefore be implemented that provides preferential attachment between layers based on Q_α. The following growth mechanisms are proposed.

The network is grown by layer, with a new node, *i*, introduced every timestep on every layer with core-discipline D_i connecting to m_0 previously active node entities, *j*, on layer D_i. The node entities are chosen based on their degree given by $\Phi_{\mathrm{i}}^{\alpha}$.

$$\Phi_{\mathrm{i,t}}^{\alpha} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \tag{9.59}$$

This represents that individuals who have more prominence are more likely to attract new collaborators.

At every timestep, every node entity has a probability, $\Psi_{\mathrm{i}}^{\alpha}$, to link to m_1 previously active node entities. To mimic the preferential attachment process, $\Psi_{\mathrm{i}}^{\alpha}$ is given by the following expression.

$$\Psi_{\mathrm{i,t}}^{\alpha} = \Phi_{\mathrm{i,t}}^{\alpha} C_0 N_t^{\alpha} \tag{9.60}$$

This represents that individuals who have more prominence are more likely to create new connections as they are more active researchers in that discipline.

This can then attach itself to other node entities in the same or other layers. However, as preferential attachment cannot link a node to itself (and the desirable structure does not occur if it does), two separate probability mechanisms are proposed for connecting to node entities within the layer and outside the layer. These are given in the following expression.

$$\Theta_{\mathrm{j,t}}^{\alpha\beta} = \begin{cases} q\Phi_{\mathrm{i,t}}^{\alpha}, & if\ \alpha = \beta \\ \dfrac{1}{N_t^{\beta}}(1-q)\dfrac{{Q_{\beta}}^{\gamma}}{\sum_{\mu}^{M}({Q_{\mu}}^{\gamma})}, & if\ \alpha \neq \beta \end{cases} \tag{9.61}$$

Where γ is a tuning parameter and q is a value between 0 and 1, which determines the proportion of m_1 links that are in layer α, when the originating node entity ($\Psi_{\mathrm{i,t}}^{\alpha}$) also occurs in layer α.

When a new collaboration occurs between node entities i and j on layers α and β respectively, a link is created between i and j on layers α, D_i, and D_j.

This represents that if someone else attracts a new collaboration, it is more likely to be individuals who have more prominence if it is in the same field, or to an individual who is active in a more interdisciplinary field if it is IDR. However, the link is formed in the core-disciplines of the respective researchers and the field from which it originated (D_i, D_j, and α respectively).

In layman's terms, this model adds one new node per layer, which connects to m_0 existing node entities on that layer based on those node entities' degrees ($\Phi_{\mathrm{i,t}}^{\alpha}$). At the same time, every node entity in the layer has a probability based on its degree ($\Psi_{\mathrm{i,t}}^{\alpha}$) to connect to m_1 node entities. These target node entities are chosen based on which layer they are on. The node entities in the same layer have probability $\Theta_{\mathrm{j,t}}^{\alpha\alpha}$, while node entities outside the layer have probability $\Theta_{\mathrm{j,t}}^{\alpha\beta}$.

This mimics the following.

- New node entities preferring highly connected node entities in the same discipline when collaborating.
- Existing node entities collaborating more if they are highly connected.
- Highly connected node entities being preferable to collaborate within the same discipline.
- Interdisciplinary collaborations occur more frequently in disciplines with many previous interdisciplinary collaborations, and nodes with a large number of interdisciplinary activity (modelled implicitly with node entities).

9.7.1. Degree distributions

This section establishes whether the Model 4's degree distribution matches the real-world network's degree distribution.

Hypothesis 9.1: The degree distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.

The simulation differs from the pure Barabási-Albert model (Model 1) in that the degree coefficient fits well, but with a lower exponent. It compares very well to the University of Bath multiplex co-authorship network's aggregate degree distribution. The exponent is roughly ~-2.25, well within the range given to corroborate Hypothesis 9.1(a).

Hypothesis 9.1(a) - The aggregate degree distribution produces a power-law relationship with an exponent between -1.5 to -2.5

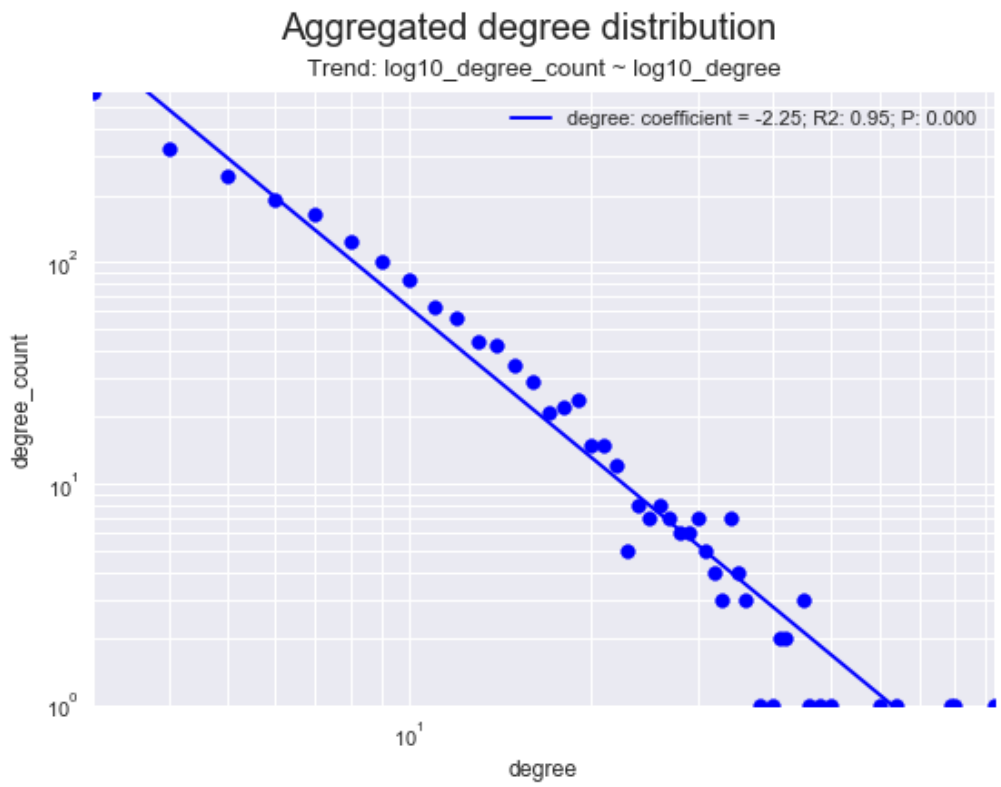

Figure 9.54. Aggregate degree distribution for Model 4.

All layers produce statistically significant degree exhibiting a strong power-law relationship. This is seen in all, disciplinary, and interdisciplinary node entities as seen in Figure 9.55 to Figure 9.57. As such, Hypothesis 9.1(b) is corroborated.

Hypothesis 9.1(b) - The degree distribution on every layer produces a power-law relationship using all node entities, disciplinary node entities only, and interdisciplinary node entities only.

A few differences between the disciplinary and interdisciplinary node entities can be seen. The disciplinary node entities show relatively few low degree nodes and many high degree nodes in

comparison to its interdisciplinary counterpart. This results in the interdisciplinary node entities' degree distributions having a much larger exponent magnitude.

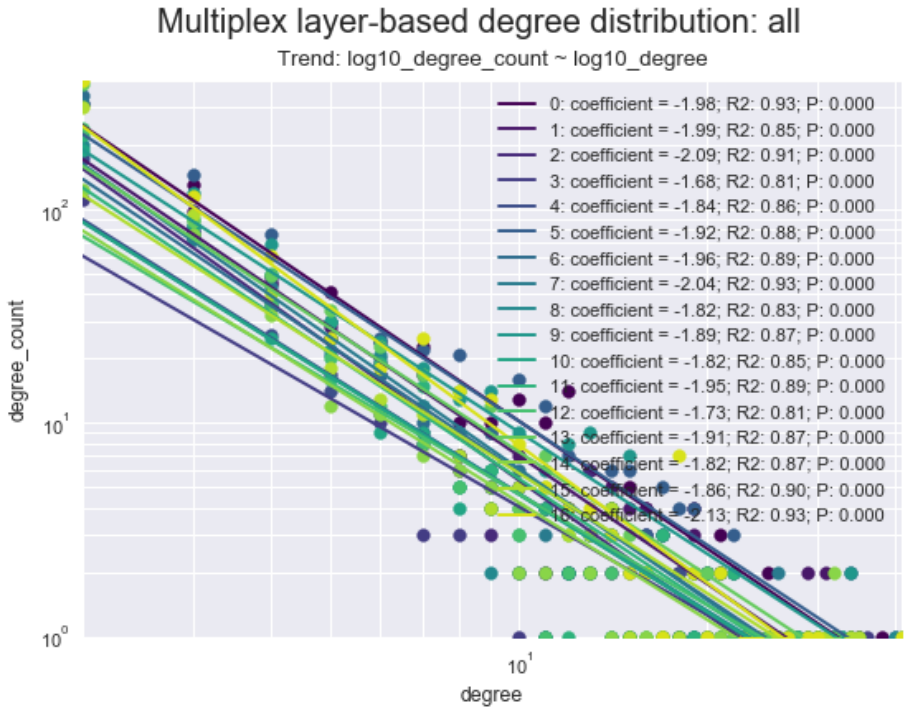

Figure 9.55. Layer degree distributions for Model 4 for all node entities.

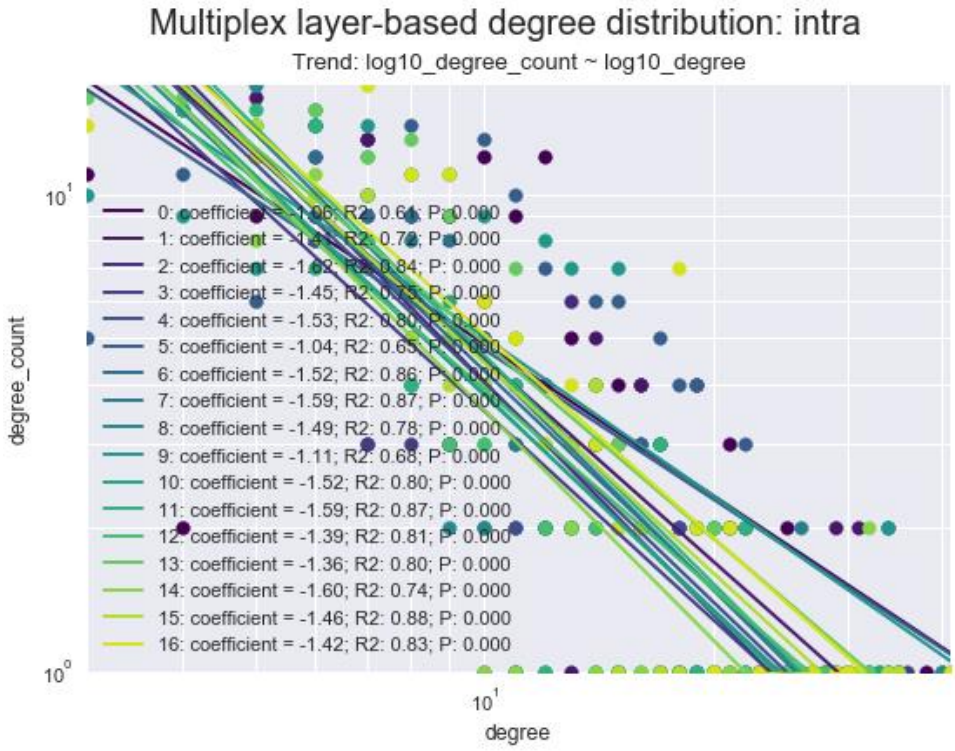

Figure 9.56. Layer degree distributions for Model 4 for disciplinary node entities only.

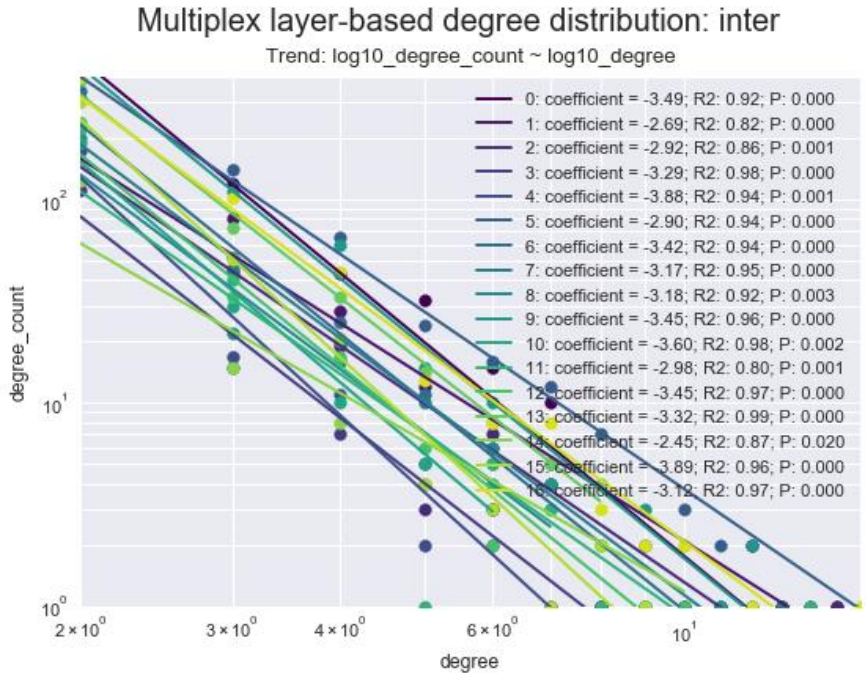

Figure 9.57. Layer degree distributions for Model 4 for interdisciplinary node entities only.

Figure 9.58 show the distribution of the exponents. The exponents form a Gaussian distribution, although ~22% (across all runs) of layers show disciplinary degree distributions having exponents above -1.40, suggesting that disciplinary node entities may be skewed left.

Despite the distribution of exponents not matching perfectly (mostly unskewed) compared to the real-world networks (disciplinary node entities skewed right), the exponent trends match Hypotheses 9.1(c)-(e), which are corroborated, whist Hypothesis 9.1(f) is rejected.

Hypothesis 9.1(c) - The degree distribution on every layer, using every node entity, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than the aggregate exponent.

Hypothesis 9.1(d) - The degree distribution on every layer, using disciplinary node entities only, produces power-law exponents whose peak KDE density occurs at an exponent slightly lower than all the node entities' peak exponent.

Hypothesis 9.1(e) - The degree distribution on every layer, using interdisciplinary node entities only, produces power-law

exponents whose peak KDE density occurs at an exponent above the aggregate exponent.

Hypothesis 9.1(f) - The degree distributions' exponents are distributed as Gaussians that are skewed to the right as estimated by the KDE.

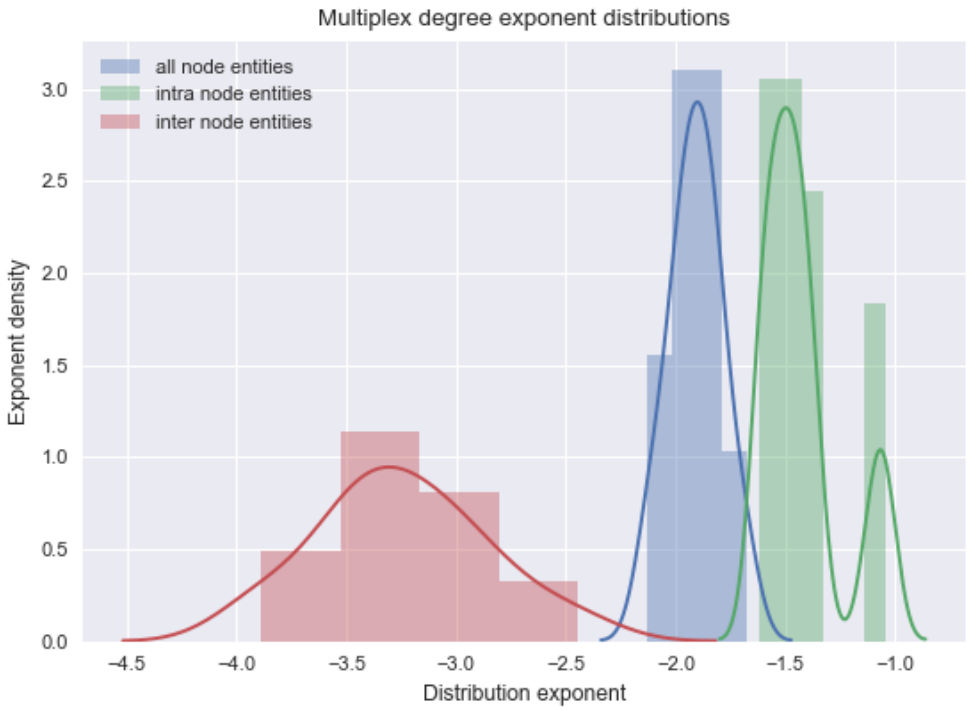

Figure 9.58. Layer power-law exponent distributions for Model 4.

The dynamics have not changed much from the previous model: the network is growing and gives advantage to early nodes, preferential attachment happening within the layer according to $\Phi^{\alpha}_{i,t}$, random connections occurring for node entities with probability $\Psi^{\alpha}_{i,t}$, and a probability to connect to all nodes' core-discipline node entities $\Theta^{\alpha}_{i,t}$.

Two of these are unchanged from the previous model with only $\Psi^{\alpha}_{i,t}$, and $\Theta^{\alpha}_{i,t}$ having different values. $\Psi^{\alpha}_{i,t}$ and $\Theta^{\alpha\alpha}_{i,t}$ will favour highly connected nodes across all layers in both disciplinary and interdisciplinary connections. $\Theta^{\alpha\beta}_{i,t}$ will favour layers only, whilst node entities are chosen at random. However, it is mainly through $\Theta^{\alpha\beta}_{i,t}$ that interdisciplinary connections occur (unless there are more interdisciplinary node entities than disciplinary node entities, this is not the case). This means that this property should have no major impact on the interdisciplinary degree compared to the previous model.

As before, new nodes are guaranteed a starting disciplinary node entity degree of $k_i^{\alpha}(t_i) = m_0$, whereas interdisciplinary node entities will only ever start with $k_i^{\alpha}(t_i) = 1$. The growth mechanism ensures that there will be more high degree nodes that are disciplinary. This is one of the mechanisms driving power-law relationships.

Contribution to knowledge

A central aspect of the growth models is that preferential attachment occurs (Albert and Barabási 2002), but that it occurs separately for every discipline. This means that research is attracted to individuals who have more collaborative exposure (i.e. highly connected nodes). Disciplinary authors have an advantage. A disciplinary author is more likely to collaborate with their peers in the same discipline just by virtue of them being present there. An interdisciplinary person having less exposure to the discipline needs to collaborate more to gain a similar exposure. For IDR researchers without such exposure, randomly assigned links based on how interdisciplinary a discipline is the only recourse for gaining such exposure in the first place.

This implies that for individuals to enable IDR, two aspects need to be considered: the activity of the node and the interdisciplinarity of the discipline in which they are active in.

To sustain IDR, the prominence in the discipline of interest is key.

Hypothesis 9.1 is mostly corroborated, but the correct skewness has not been achieved.

9.7.2. Disciplinary vs interdisciplinary degree regression.

The disciplinary versus sum of interdisciplinary degrees show that the model has been calibrated to be approximately equal to the University of Bath multiplex co-authorship network. Therefore, Hypothesis 9.2 is corroborated.

> ***Hypothesis 9.2: The disciplinary node entities degrees are larger than the median of the sum of their counterpart interdisciplinary node entities' degrees.***

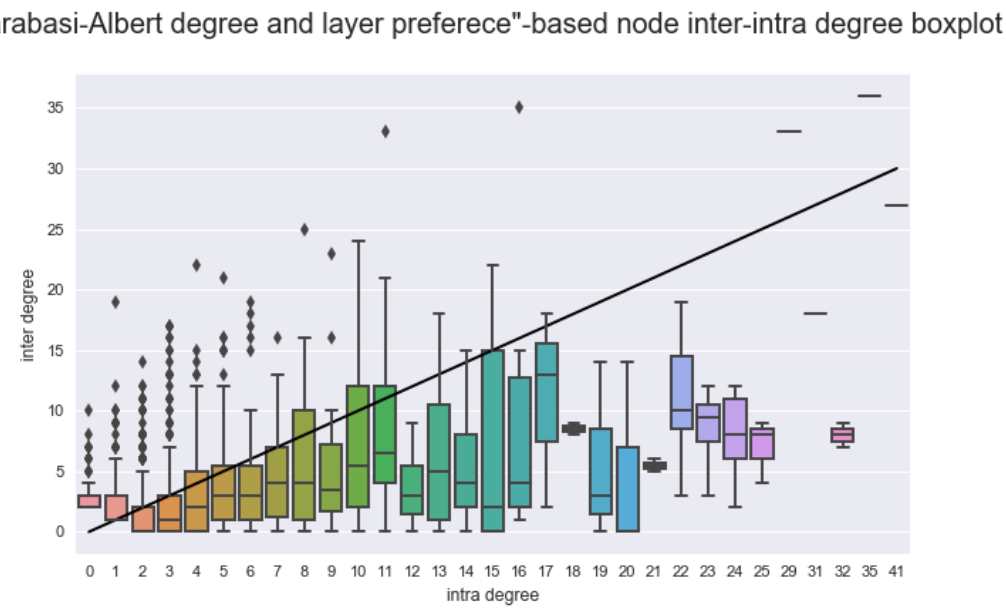

Figure 9.59. Layer power-law exponent distributions for Model 4.

This model also provides a linear trend, suggesting that the barriers to IDR have been captured.

9.7.3. Degree-correlations

This section establishes whether the degree-correlations in this model match the degree-correlations of the real-world network.

> ***Hypothesis 9.3: The degree-correlation distribution of the network matches the degree distribution of the University of Bath multiplex co-authorship network.***

The aggregate degree-correlation still exhibits a weak negative trend as can be seen in Figure 9.60. This does not show the slightly positive trend in the real-world network.

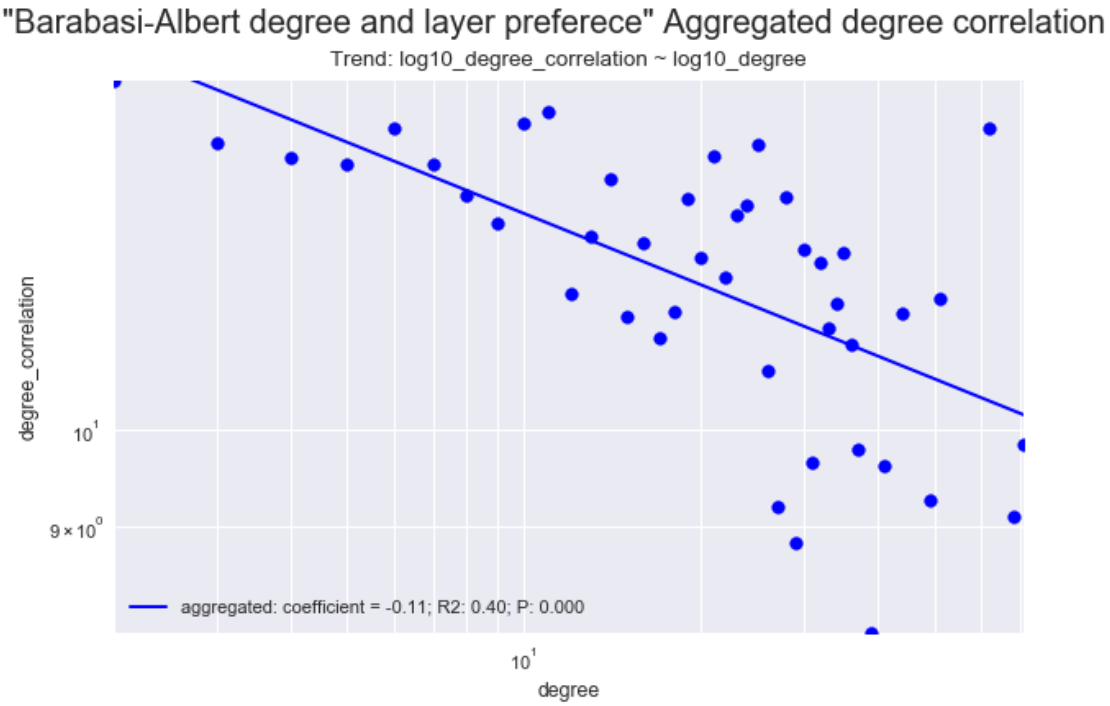

Figure 9.60. The degree-correlation for Model 4's aggregate network.

This trend is matched by the layer results, and is largely driven by the disciplinary nodes, which explains the slightly negative trend. Hypothesis 9.3(a) is therefore corroborated.

Hypothesis 9.3(a) - Layers exhibit degree-correlation distributions with a power-law relationship with a negative exponent.

This is matched on the individual layers with minimal deviation as shown in Figure 9.61. The disciplinary node entities are driving this, as it matches all node entities very closely. In the interdisciplinary node entities, there is a significant amount of variation, and statistically insignificant results, resulting in the wider spread seen. However, of the statistically significant results, the interdisciplinary node entities' degree-correlations peak density occurs at ~-0.44. This result is questionable as the statistically insignificant results have smaller exponents, which could skew the exponents' distribution above -0.3. As it is, Hypothesis 9.3(b) is only partially corroborated.

Hypothesis 9.3(b) - Degree-correlation distribution exponents exhibit Gaussian distributions as estimated by the KDE skewed right with the peak density occurring at a value of $\gamma > -0.3$.

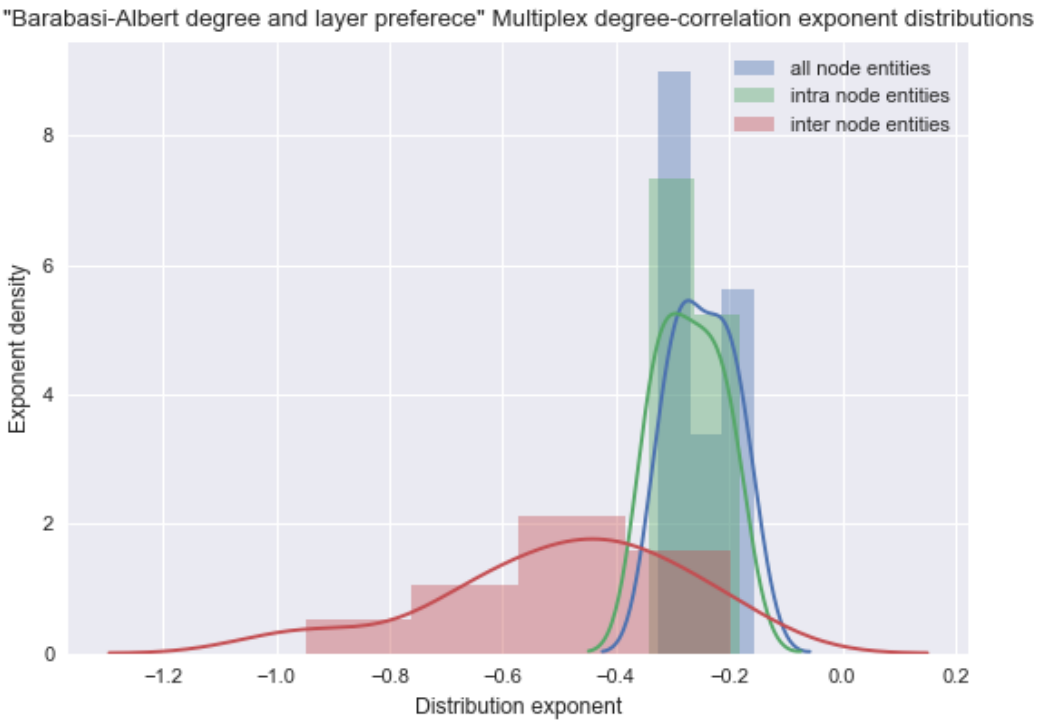

Figure 9.61. The degree-correlation exponents distribution for Model 4.

9.7.4. Node activity

The node activity produces a similar distribution as to when the links were added to random node entities. This suggests that the node activity is a key driver to the selection and creates a node activity distribution as seen in Figure 9.62.

The power-law exponent is similar to the real-world results (-2.67 compared to -3.32).

The biggest difference is the slight curve in the simulation results. Whilst a power-law relationship is still statistically significant, the bend occurs at nine layers of activity.

However, given that the trend is statistically significant, Hypothesis 9.4 is corroborated.

> ***Hypothesis 9.4: The multiplex node activity exhibits a power-law distribution with a negative exponent between*** $-2.5 \geq \gamma \geq -3.5$.

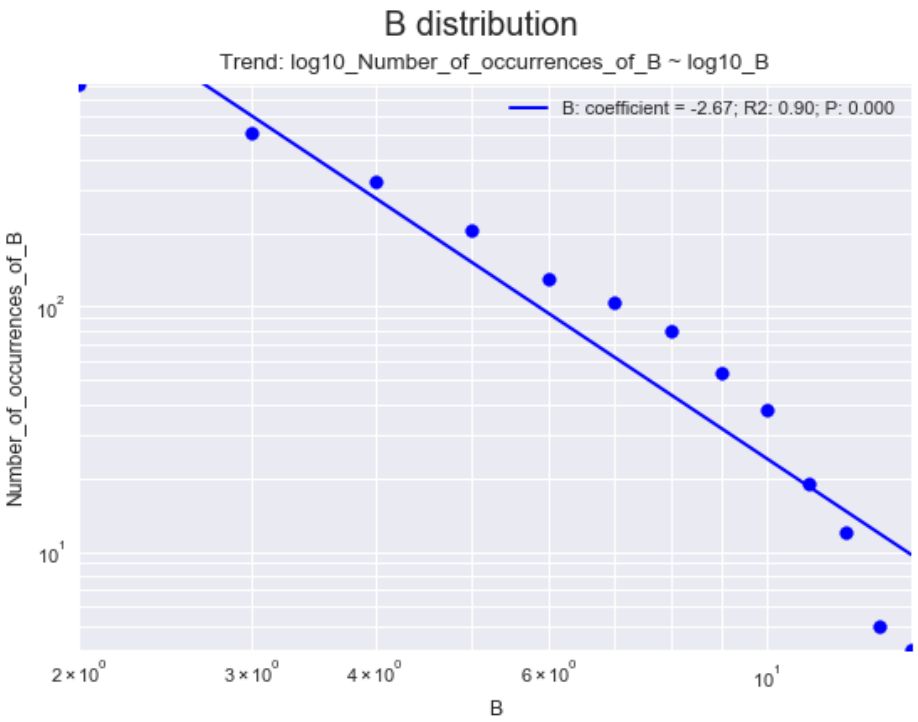

Figure 9.62. Node activity distribution for Model 4.

The node activity provides the same preferential attachment mechanism ensuring that a power-law distribution occurs. This implies that individuals who are more interdisciplinary are more like to become even more interdisciplinary, and are therefore better suited to enable IDR.

9.7.5. Layer activity

As there are few layers, the sample size is too small to create a statistically significant distribution as can be seen in Figure 9.63. However, a linear relationship is not completely inappropriate. The p-value ranges around ~0.25, suggesting that this distribution has a 25% of occurring if it was not a linear distribution. The exponents are very similar to the real-world network's results (-0.02 compared to -0.01).

Given the nature of $\Theta_{i,t}^{\alpha\beta}$ a power-law distribution was expected, as it provides a layer advantage for those with high Q^{α}.

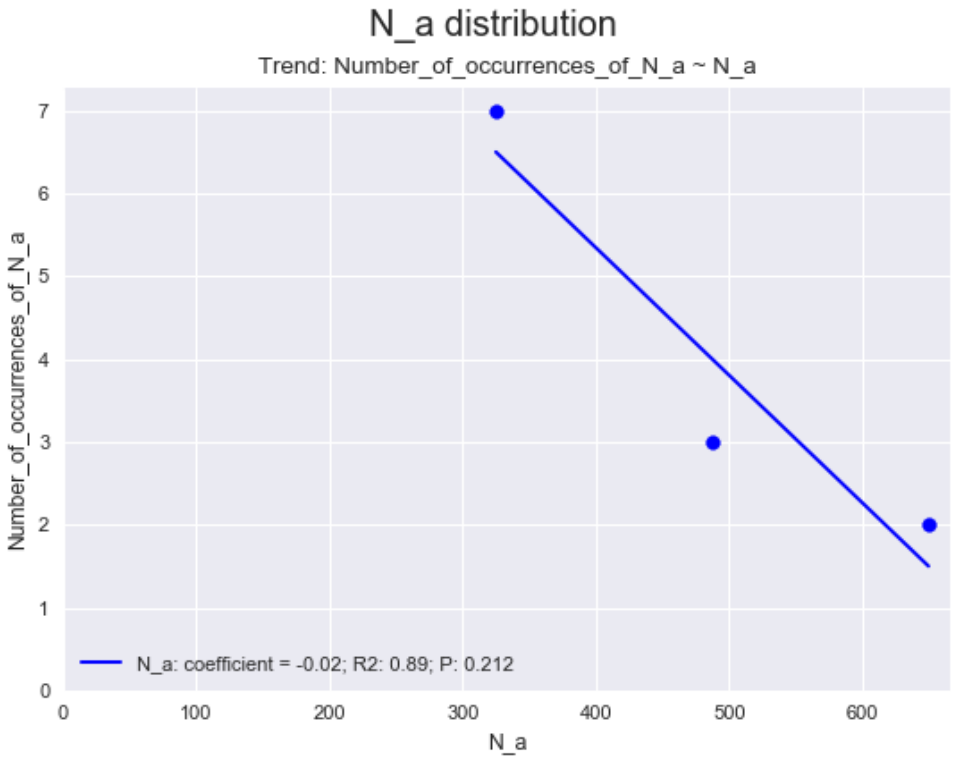

Figure 9.63. Node activity distribution for Model 4.

Given that the growth of core-discipline node entities the networks are occurring at a given rate, this distribution must occur due to IDR. It therefore stands to reason that this approximation occurs due to heterogeneous attractiveness of conducting in one discipline over another. Ultimately, this has to be due to the discipline layer closeness parameter. This means that disciplines with high amount of IDR attract more IDR either due to the interdisciplinarity of the discipline's knowledge, due to the central location of the discipline, or some combination of these.

9.7.6. Layer-pair closeness

The layer-pair closeness is arguably the most important measure for this model as it represents the most significant measure that has not been achieved by other models.

Figure 9.64 shows the histogram and KDE plot of the layer-pair closeness. A fat-tail has clearly been created, with the peak occurring at the smallest value.

Figure 9.65 shows that a power-law with an exponent of -0.97 is statistically significant. This compares well to the University of Bath multiplex co-authorship network which exhibits an exponent of -0.65.

An interesting aspect is that few rules allow for this distribution to occur in the simulation. For instance, rules using $Q_{\alpha\beta}$ as a preferential attachment produce Gaussian distributed $Q_{\alpha\beta}$.

Given that the results are statistically significant, Hypothesis 9.5 is corroborated.

Hypothesis 9.5: The multiplex layer-pair closeness exhibits a power-law distribution with a negative exponent between $-0.3 \geq \gamma \geq -1.0$.

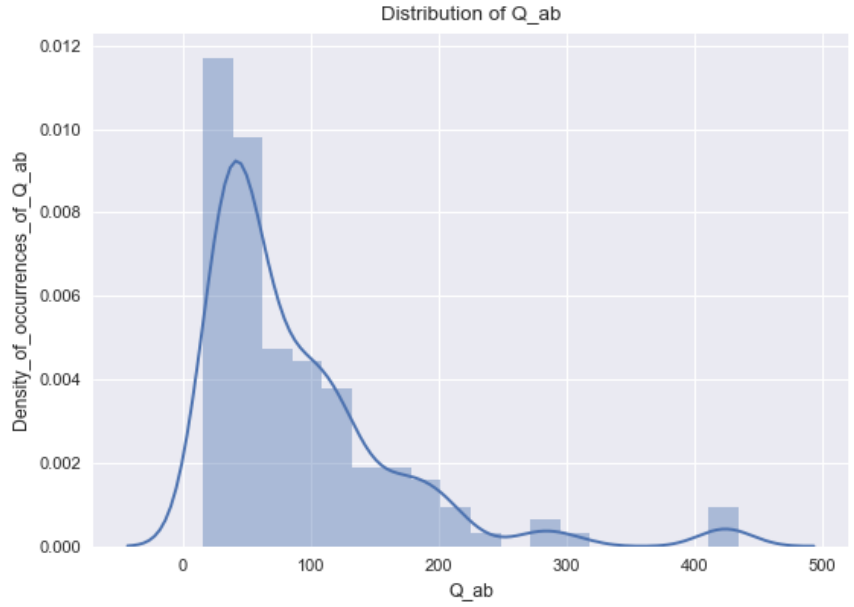

Figure 9.64. Layer-pair closeness distribution for Model 4 as a histogram and KDE plot.

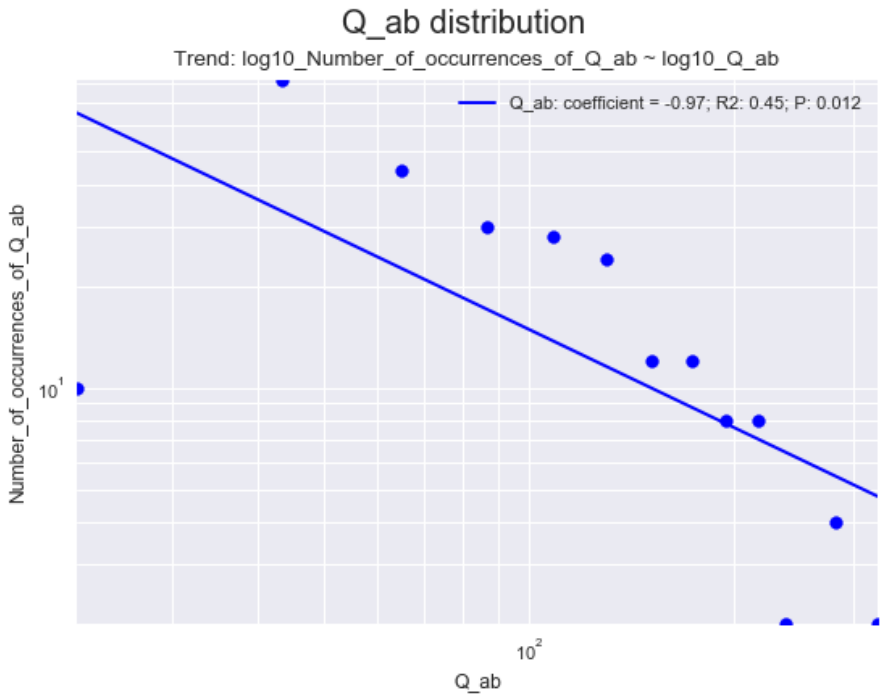

Figure 9.65. Layer-pair closeness for Model 4 as a distribution with a power-law relationship.

It is important to note that this distribution establishes layer-pair closeness. That is to say that it would imply that there is some naturally occurring mechanism that allows for two disciplines to become more interdisciplinary over time. This could for instance be a knowledge diffusion and creation issue, where the more a discipline collaborates with others, the more readily it is able to share or accept its research paradigms, administration, language, and all the IDR inhibitors discussed in Chapter 3 with that other discipline.

However, given that there is no mechanism that suggests this. The analytical analysis (see section 9.7.7) shows that it is the layer-pair closeness centrality, Q_α, that drives this.

The Q_α distribution follows a power-law, which is found in both the simulation and the real-world results with very similar values.

The model, which is able to recreate most major structural properties finds that the interdisciplinarity is the most important factor. This supports the heatmap of the University of Bath layer-pair closeness values (see Figure 9.15), which shows that the most interdisciplinary pairs belong to the most interdisciplinary discipline (Chemistry). This is a strong indication that centrality, resources, and previous interdisciplinary experience are the biggest drivers for discipline-wide IDR.

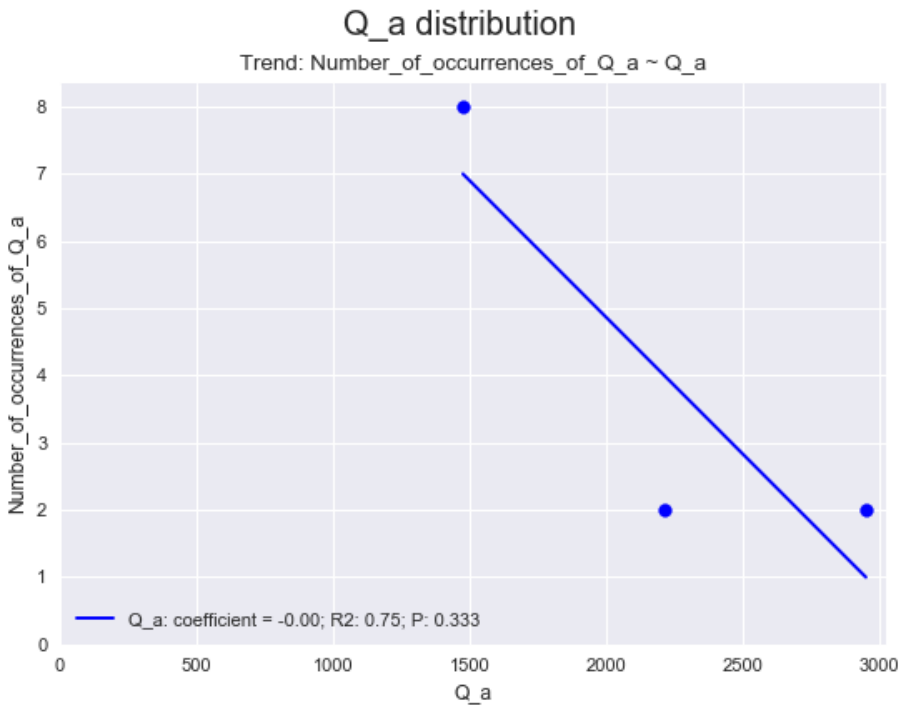

Figure 9.66. Layer closeness centrality for Model 4.

This model therefore captures the 'vertical' structure-wide mechanics. The implication is that there is also evidence that fields become more interdisciplinary over time. This could be either due to the inherent progression of a discipline's interdisciplinarity, due to the central position in the overall collaboration network, or some combination of these two.

9.7.7. Analytical analysis

As was outlined in Model 3, developing analytical expressions for the rate of change of the degree, node activity, and layer-closeness can provide valuable insights. However, it also provides a mathematical model for how these properties increase. The rate of change of the degree would therefore provide a model of who it is that can enable and sustain IDR.

The following expression recaps the components of Model 4.

$$\Phi_{\mathrm{i,t}}^{\alpha} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \tag{9.62}$$

$$\Psi_{\mathrm{i,t}}^{\alpha} = \Phi_{\mathrm{i,t}}^{\alpha} C_0 N_t^{\alpha} \tag{9.63}$$

$$\Theta_{\mathrm{j,t}}^{\beta\alpha} = \begin{cases} Mq\Phi_{\mathrm{i,t}}^{\alpha}, & if\ \alpha = \beta \\ \dfrac{1}{N_t^{\beta}} M(1-q) \dfrac{{Q_{\beta}}^{\gamma}}{\sum_{\mu}^{M}\left({Q_{\mu}}^{\gamma}\right)}, & if\ \alpha \neq \beta \end{cases} \tag{9.64}$$

9.7.7.1. Rate of change of node-entity degree

To develop an expression for the rate of change of degree, it is necessary to realise that one must develop an expression for $\frac{dk_i^{\alpha=D_i}}{dt}$ and $\frac{dk_i^{\alpha\neq D_i}}{dt}$. There are three mechanisms by which a node can increase its degree:

- By connecting a new node.
- By creating a new link to an old node.
- By receiving a new link from an old node.

This results in the following expressions.

$$\frac{dk_i^{\alpha=D_i}}{dt} = m_0\Phi_{\mathrm{i,t}}^{\alpha} + B_i m_1 \Psi_{\mathrm{i,t}}^{\alpha} + m_1 \sum_{j=1}^{N_t^{\alpha}} \left(\Psi_{j,\mathrm{t}}^{\alpha}\right) \cdot \Theta_{i,\mathrm{t}}^{\alpha\alpha} + m_1 \sum_{\beta\neq\alpha}^{M-1} \left(\sum_{j=1}^{N_t^{\beta}} \left(\Psi_{j,\mathrm{t}}^{\beta}\right) \right) \cdot (B_i - 1) \sum_{\beta\neq\alpha}^{M-1} \left(\Theta_{i,\mathrm{t}}^{\alpha\beta}\right) \tag{9.65}$$

$$\frac{dk_i^{\alpha\neq D_i}}{dt} = m_0\Phi_{\mathrm{i,t}}^{\alpha} + m_1\Psi_{\mathrm{i,t}}^{\alpha} + m_1 \sum_{j=1, i\neq j}^{N_t^{\alpha}} \left(\Psi_{j,\mathrm{t}}^{\alpha}\right) \cdot \Theta_{i,\mathrm{t}}^{\alpha\alpha} + m_1 \Psi_{i,\mathrm{t}}^{D_i} \Theta_{i,\mathrm{t}}^{D_i\alpha} \tag{9.66}$$

Disciplinary node entities

To establish the mechanics of disciplinary node entities, it is necessary to solve for equation 9.65.

$$\begin{aligned}\frac{dk_i^{\alpha=D_i}}{dt} = \; & m_0\Phi_{\mathrm{i,t}}^{\alpha} \\ & +B_i m_1 \Phi_{\mathrm{i,t}}^{\alpha} C_0 N_t^{\alpha} \\ & +m_1 \sum_{j=1}^{N_t^{\alpha}} \left(\Phi_{\mathrm{j,t}}^{\alpha} C_0 N_t^{\alpha}\right) \cdot Mq\Phi_{\mathrm{i,t}}^{\alpha} \\ & +m_1 \sum_{\beta\neq\alpha}^{M-1} \left(\sum_{j=1}^{N_t^{\beta}} \left(\Phi_{j,\mathrm{t}}^{\beta} C_0 N_t^{\beta}\right) \right) \\ & \cdot (B_i - 1) \sum_{\beta\neq\alpha}^{M-1} \left(\frac{1}{N_t^{\beta}} M(1-q) \frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \right)\end{aligned} \tag{9.67}$$

As $\sum_{j=1}^{N_t^{\alpha}}\left(\Psi_{j,\mathrm{t}}^{\alpha}\right) = C_0 N_t^{\alpha}$, this can be written as follows.

$$\begin{aligned}\frac{dk_i^{\alpha=D_i}}{dt} = & \; m_0\Phi_{\mathrm{i,t}}^{\alpha} \\ & +B_i m_1 \Phi_{\mathrm{i,t}}^{\alpha} C_0 N_t^{\alpha} \\ & +qm_1 C_0 N_t^{\alpha} \cdot \Phi_{\mathrm{i,t}}^{\alpha} \\ & +(1-q)m_1 C_0 (M-1)\left\langle N_t^{\beta}\right\rangle (B_i - 1)(M-1)\left\langle \frac{1}{N_t^{\beta}} \frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \right\rangle^{\beta\neq D_i}\end{aligned} \tag{9.68}$$

$$\begin{aligned}\frac{dk_i^{\alpha=D_i}}{dt} = & \; \Phi_{\mathrm{i,t}}^{\alpha}\left(m_0 + m_1 C_0 N_t^{\alpha}(B_i + Mq)\right) \\ & +M(1-q)(M-1)^2 m_1 C_0 \left\langle N_t^{\beta}\right\rangle^{\beta\neq D_i} \left\langle \frac{1}{N_t^{\beta}} \frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \right\rangle^{\beta\neq D_i} \cdot (B_i - 1)\end{aligned} \tag{9.69}$$

$$\frac{dk_i^{\alpha=D_i}}{dt} = A_0(B_i + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1 A_2 \left\langle \frac{1}{N_t^{\beta}} \frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \right\rangle^{\beta\neq D_i} \cdot (B_i - 1) \tag{9.70}$$

Where:

- $A_0 = (m_0 + m_1 C_0 N_t^{\alpha})$, a constant for discipline α
- $A_1 = M(1-q)m_1 C_0$, a constant for discipline α
- Assuming $\left\langle N_t^{\beta}\right\rangle^{\beta\neq D_i} \sim \left\langle N_t^{\beta}\right\rangle$

- $A_2 = \langle N_t^{\beta} \rangle (M-1)^2$, a constant for discipline α
- $\frac{N_{t+1}^{\alpha} - N_t^{\alpha}}{N_t^{\alpha}} \approx 0$

As A_0 affects all nodes on layer α, the scale-free property will emerge in this model with the first term tending towards $P(k) \sim k^{-3}$.

As can be seen, the disciplinary node entities' rate of change of degree depends on two terms:

- $(B_i + Mq) \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}}$
- $(B_i - 1) \langle \frac{1}{N_t^{\beta}} \frac{{Q_{\beta}}^{\gamma}}{\sum_{\mu}^{M} ({Q_{\mu}}^{\gamma})} \rangle^{\beta \neq D_i}$

This means that there are three important aspects to the growth of disciplinary node entities: their prominence in their discipline, their interdisciplinarity, and how interdisciplinary their surrounding disciplines are.

The most sensitive of these terms is the prominence. This is not do with absolute value, but rather how different k_i^{α} can be in comparison to $\langle \frac{1}{N_t^{\beta}} \frac{{Q_{\beta}}^{\gamma}}{\sum_{\mu}^{M} ({Q_{\mu}}^{\gamma})} \rangle^{\beta \neq D_i}$. The second most important is the interdisciplinarity of the researcher, B_i. However, all three play an important role in the growth of a researcher's collaboration network.

Interdisciplinary node entities

To establish the mechanics of interdisciplinary node entities, it is necessary to solve for equation 9.66. Using similar assumptions as above, the following expression can be obtained.

$$\frac{dk_i^{\alpha \neq D_i}}{dt} = A_0(1+Mq) \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1 \frac{N_t^{D_i}}{N_t^{\alpha}} \frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}} \frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M} ({Q_{\mu}}^{\gamma})} \qquad (9.71)$$

As can be seen, the disciplinary node entities' rate of change of degree depends on two terms:

- $\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}}$
- $\frac{N_t^{D_i}}{N_t^{\alpha}} \frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}} \frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M} ({Q_{\mu}}^{\gamma})}$

The first term is simply the prominence of the researcher in the discipline, and is actually the largest term as $A_0(1+Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \gg A_1 \frac{N_t^{D_i}}{N_t^{\alpha}} \frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}} \frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M}({Q_{\mu}}^{\gamma})}$.

This means that when identifying the individuals who sustain IDR, it is vital to identify the IDR researcher with the greatest prominence in the specific discipline. This also implies that there is no significant quality that makes one IDR researcher great in all IDR research, but rather that the researcher is proficient between two different disciplines.

The second term is important too, and consists of three different components:

- The ratio of how many researchers there are in the originating discipline compared to the target discipline.
- The prominence of the IDR researcher in their own discipline.
- The interdisciplinarity of the target discipline.

This second term is the term that identifies individuals who are most likely to break down the barriers of IDR in the absence of any presence, and is therefore vital. Without this term, no IDR could occur.

Assuming that every term is directly applicable to a real-world phenomenon, it is useful to reflect on what each component could represent:

- The ratio could be indicative of the originating discipline's resources in comparison to the target discipline's. Therefore, with more resources, it is more likely that an outward IDR will occur.

  In a direct comparison to the layer-pair closeness as outlined in the University of Bath multiplex structure (see section 9.3.1.6), Chemistry is central to the University of Bath, and collaborates a lot with other disciplines. This ratio could be reminiscent of such a dynamic, but would suggest that two equally sized disciplines would be less likely to collaborate in any one direction.
- The prominence of the IDR researcher in their own discipline could be indicative of ability. It could equally be indicative of their attractiveness to collaborate with.
- The interdisciplinarity of the target discipline could be indicative of reduced barriers to conducting IDR in that discipline.

Therefore, the second term can be seen as enabling IDR, and the first term can be seen as enabling and sustaining IDR. Decision and policy makers should therefore target individuals who have prominence in the target discipline first, then the prominence in their originating discipline second, and then consider how easy it would be to collaborate in the target discipline.

Contribution to knowledge:

A model has been created based on growing a network that adds links preferentially within a discipline and based on interdisciplinarity between disciplines. By virtue of nodes having an interdisciplinary presence within other disciplines, they can take advantage of the former mechanism.

Such a model accurately mimics the University of Bath 2000-2017 network structure. As the growth model has been made explicit it is possible to create an expression for the rate of change of number of collaborators. The following expressions express the rate of change for disciplinary and interdisciplinary node entities respectively.

$$\frac{dk_i^{\alpha=D_i}}{dt} = A_0(B_i + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1A_2(B_i - 1)\left\langle \frac{1}{N_t^{\beta}} \frac{{Q_{\beta}}^{\gamma}}{\sum_{\mu}^{M}\left({Q_{\mu}}^{\gamma}\right)} \right\rangle^{\beta \neq D_i} \tag{9.72}$$

$$\frac{dk_i^{\alpha\neq D_i}}{dt} = A_0(1 + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1\frac{N_t^{D_i}}{N_t^{\alpha}}\frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}}\frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M}\left({Q_{\mu}}^{\gamma}\right)} \tag{9.73}$$

Where A_0, A_1, and A_2 are constants, k_i^{α} is the degree of node i in discipline α, ${Q_{\beta}}^{\gamma}$ is the interdisciplinarity of discipline β to the power of γ, B_i is the node activity of node i, q is a (probability) measure of how disciplinary new links are, and M is the number of disciplines. This expression represents the overall model for identifying not just interdisciplinary leaders of tomorrow, but all leaders in specific layers. A rate of change measure is a linear extrapolation, it is therefore a predictive measure. Therefore, individuals who have a high rate of change of degree enable and sustain IDR. This therefore represents a mathematical model that the research aim set out to achieve. This represents a significant contribution to knowledge.

9.7.7.2. Rate of change of node activity

Given the importance of node activity in the overall model, it is useful to develop an expression for node activity. However, as there are many interlayer dependencies a generalised expression for the actual node activity would be unwieldy. It is therefore useful to consider the rate of change of node activity.

The node activity is affected by both changes to the rules: $\Psi_{i,t}^{\alpha}$ and $\Theta_{j,t}^{\alpha\beta}$. The rate of change of node activity in large networks dominated by disciplinary nodes in every layer is given by the following expressions.

$$\frac{B_i}{dt} = B_i\left(m_1\Psi_{\mathrm{i,t}}^{\alpha}\sum_{\beta\neq\alpha}^{M}\Theta_{j,\mathrm{t}}^{\alpha\beta} + m_1\Theta_{i,\mathrm{t}}^{\beta\alpha}\sum_{\beta\neq\alpha}^{M}\sum_{j\neq i}^{N^{\beta}}\Psi_{\mathrm{j,t}}^{\beta}\right) \tag{9.74}$$

$$\frac{dB_i}{dt} \cong B_i\left(m_1(M-1)\Psi_{\mathrm{i,t}}^{\alpha}\langle\Theta_{j,\mathrm{t}}^{\alpha\beta}\rangle + m_1(M-1)\left(N^{\beta}-1\right)\Theta_{i,\mathrm{t}}^{\beta\alpha}\langle\Psi_{\mathrm{j,t}}^{\beta}\rangle\right) \tag{9.75}$$

$$\frac{dB_i}{dt} \cong B_i m_1 C_0(M-1)(1-q)\left(\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}}k_j^{\alpha}}\frac{N_t^{\alpha}}{\langle N_t^{\beta}M\rangle} + \frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M}\left({Q_{\mu}}^{\gamma}\right)}\right) \tag{9.76}$$

The scale-free in the randomly added links was driven by B_i. In this simulation it is driven by $B_i\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}}k_j^{\alpha}}$ and $B_i\frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M}({Q_{\mu}}^{\gamma})}$, which could explain the curve that was found in the simulation.

This shows how interconnected all the various terms are and highlights how some of the findings can be emergent.

9.7.7.3. Rate of change of layer-pair closeness

Given that the developed expression for interdisciplinary node entities highlighted the importance of originating and target disciplines, it is important to gain some insight into how it is that layer-pairs develop.

The layer pair closeness is approximated by the following expression.

$$\frac{dQ_{\alpha\beta}}{dt} = m_1\sum_{j\neq i}^{N^{\alpha}}\Psi_{\mathrm{i,t}}^{\alpha}\sum_{j\neq i}^{N^{\beta}}\Theta_{j,\mathrm{t}}^{\alpha\beta} + m_1\sum_{j\neq i}^{N^{\alpha}}\Theta_{i,\mathrm{t}}^{\beta\alpha}\sum_{j\neq i}^{N^{\beta}}\Psi_{\mathrm{j,t}}^{\beta} \tag{9.77}$$

$$\frac{dQ_{\alpha\beta}}{dt} = m_1C_0(1-q)\left(N_t^{\alpha}\langle\frac{k_i^{\alpha}}{\langle k_j^{\alpha}\rangle}\rangle\langle\frac{{Q_{\beta}}^{\gamma}}{M\langle{Q_{\mu}}^{\gamma}\rangle}\rangle + N_t^{\beta}\langle\frac{{Q_{\alpha}}^{\gamma}}{M\langle{Q_{\mu}}^{\gamma}\rangle}\rangle\langle\frac{k_j^{\beta}}{\langle k_m^{\beta}\rangle}\rangle\right) \tag{9.78}$$

Where $\langle\frac{k_i^{\alpha}}{\langle k_j^{\alpha}\rangle}\rangle = \frac{1}{N_t^{\alpha}}$

$$\frac{dQ_{\alpha\beta}}{dt} = m_1C_0(1-q)\left(\frac{{Q_{\beta}}^{\gamma}}{M\langle{Q_{\mu}}^{\gamma}\rangle} + \frac{{Q_{\alpha}}^{\gamma}}{M\langle{Q_{\mu}}^{\gamma}\rangle}\right) \tag{9.79}$$

Therefore, the power-law property of $Q_{\alpha\beta}$ is dependent on ${Q_{\beta}}^{\gamma}$ and ${Q_{\alpha}}^{\gamma}$. This therefore shows that layer-pair closeness is entirely dependent on either how central a discipline is, or how interdisciplinary they are (these are for all intents and purposes mathematically equal).

This implies that two disciplines are more likely collaborate if there have been many interdisciplinary collaborations in the past in both fields equally.

Therefore, for two disciplines to collaborate, both need to focus on reducing their barriers by engaging in IDR with as many different fields as possible and with as many different individuals as possible.

9.7.8. Discussion

The model provides a very similar structure to the University of Bath department-based multiplex co-authorship network as can be seen in Table 9.11. Given the simplicity of the model, this produces a near realistic multiplex structure, and can therefore be considered to provide a successful growth model. Model 4 corroborated all hypotheses bar 9.3(b), which was only partially corroborated. However, even that hypothesis could possibly be a good structural match as the real-world network also exhibited a lot of statistically insignificant layer distributions. Given that the model corroborated on all the key metrics and was able to mimic the statistically insignificant properties as well, there is a good argument to be made that this model is validated through historical data validation.

This provides a mathematical basis to analyse the evolution of a real network as well as providing clues to important mechanics in multiplex networks. These mechanics provide original contributions to knowledge.

As was outlined in Model 3, node entities play an important role in drowning out random elements of the model. Given the analytical analysis not accounting for such a mechanism, it must be concluded that it is an emergent property that occurs with the synergy between the node entities' degrees, the number of nodes in a layer, the node activity, and the layer closeness centrality. The influence of these on each other represents important future work that can further clarify why the multiplex networks behave as they do.

The fact that disciplinary nodes in this model are m_0 times more connected than interdisciplinary node entities provides a huge benefit to disciplinary nodes and is the main driver of the difference between disciplinary and interdisciplinary node entities. This can be thought of as disciplinary node entities having more exposure to their discipline, and interdisciplinary node entities requiring a lot of collaboration to overcome its initial disadvantage. Furthermore, interdisciplinary nodes do not gain the benefit of node activity directly.

Finally, given that a realistic model has been created, it was possible to extract a mathematical expression for how fast a node entity increases its degree. As the aim of the research was to identify individuals who enable and sustain IDR, this model provides a direct measure to do so.

Table 9.11 Comparative values for the real-world department-based multiplex networks of the University of Bath and the. Barabási-Albert with edges assigned on preference to node-layer degree and layer closeness centrality.

Measure		Department-based multiplex networks		Model 4	
		Trend	**Value**	**Trend**	**Value**
Degree-distribution	**Aggregate**	$\log_{10} y \sim \log_{10} x$	-1.87	$\log_{10} y \sim \log_{10} x$	-2.44
	All node entities	$\log_{10} y \sim \log_{10} x$	-1.55	$\log_{10} y \sim \log_{10} x$	-1.94
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-1.20	$\log_{10} y \sim \log_{10} x$	-1.51
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-2.25	$\log_{10} y \sim \log_{10} x$	-3.44
Disciplinary-interdisciplinary boxplot		$y \sim x$	0.33	$y \sim x$	<1
Degree-correlation	**Aggregate**	$\log_{10} y \sim \log_{10} x$	0.05*	$\log_{10} y \sim \log_{10} x$	-0.10
	All node entities	$\log_{10} y \sim \log_{10} x$	-0.35	$\log_{10} y \sim \log_{10} x$	-0.20
	Disciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.28	$\log_{10} y \sim \log_{10} x$	-0.26
	Interdisciplinary node entities	$\log_{10} y \sim \log_{10} x$	-0.36	$\log_{10} y \sim \log_{10} x$	-0.40*
B		$\log_{10} y \sim \log_{10} x$	-3.32	$\log_{10} y \sim \log_{10} x$	-2.55
N_a		$y \sim x$	-0.67*	$y \sim x$	-0.021*
Q_ab		$\log_{10} y \sim \log_{10} x$	-0.65	$\log_{10} y \sim \log_{10} x$	-0.91
Q_a		$y \sim x$	-12.02*	$y \sim x$	-0.0046*

*Not statistically significant.

**Poor fit

9.8. Predictive Validation

This section utilises the model developed in Model 4's analytical analysis to determine whether the model developed in simulation can be used to predict the IDR leaders of the future to a greater degree than established models.

$$\frac{dk_i^{\alpha=D_i}}{dt} = A_0(B_i + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1A_2(B_i - 1)\left\langle \frac{1}{N_t^{\beta}} \frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \right\rangle^{\beta \neq D_i} \tag{9.80}$$

$$\frac{dk_i^{\alpha\neq D_i}}{dt} = A_0(1 + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1\frac{N_t^{D_i}}{N_t^{\alpha}}\frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}}\frac{{Q_\alpha}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \tag{9.81}$$

The expressions above represents the model developed using the growth models. The parameters m_0, m_1, C_0, and q that make up A_0, A_1, and A_2 will remain the same as used when developing the model. However, these are constants, and simply ensure that an appropriate ratio exists between $(B_i + Mq)\frac{k_i^\alpha}{\sum_{j=1}^{N_t^\alpha} k_j^\alpha}$ and $(B_i - 1)\left\langle \frac{1}{N_t^\beta} \frac{{Q_\beta}^\gamma}{\sum_\mu^M ({Q_\mu}^\gamma)} \right\rangle^{\beta \neq D_i}$ for disciplinary node entities, and $\frac{k_i^\alpha}{\sum_{j=1}^{N_t^\alpha} k_j^\alpha}$ and $\frac{N_t^{D_i}}{N_t^\alpha} \frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}} \frac{{Q_\alpha}^\gamma}{\sum_\mu^M ({Q_\mu}^\gamma)}$ for interdisciplinary node entities.

This will be compared to a staple measure, which has a lot of validation: the aggregate degree, k_i. To ensure a good population, both models will be tuned to the University of Bath department-based multiplex network 2000-2012. These are then compared to the layer degrees in the University of Bath department-based multiplex networks 2000-2013 to 2000-2017, thereby analysing the predictive capability from one to five years in the future.

The following hypotheses is tested for each time frame.

> ***Hypothesis 9.6: The model correlates to future discipline-specific degrees with an R^2-value distribution with a KDE peak of at least 0.5.***

> ***Hypothesis 9.7: The model correlates to future discipline-specific degrees with a KDE peak higher than aggregate degree.***

The results correlating to the 2000-2013 period for all node entities are shown in Figure 9.67 for the model and Figure 9.68 for the aggregate degree. When observing the difference between the predictions in Figure 9.68, it can clearly be seen that the model predicts very different paths for every layer, in comparison to the aggregate degree, suggesting that the differences between the different disciplines are captured.

Figure 9.69 shows the distributions of the R^2-values separated by time-period. As can be seen, the R^2-values are much higher than the aggregate degrees and ranges from over 0.90 (2000-2013, one year in the future) to 0.60 (2000-2017, five years in the future). For every time period, Hypothesis 9.6 is corroborated.

In comparison to the aggregate degree, the model far outperforms as the aggregate degree posts an R^2-value high of 0.67 in the one-year prediction.

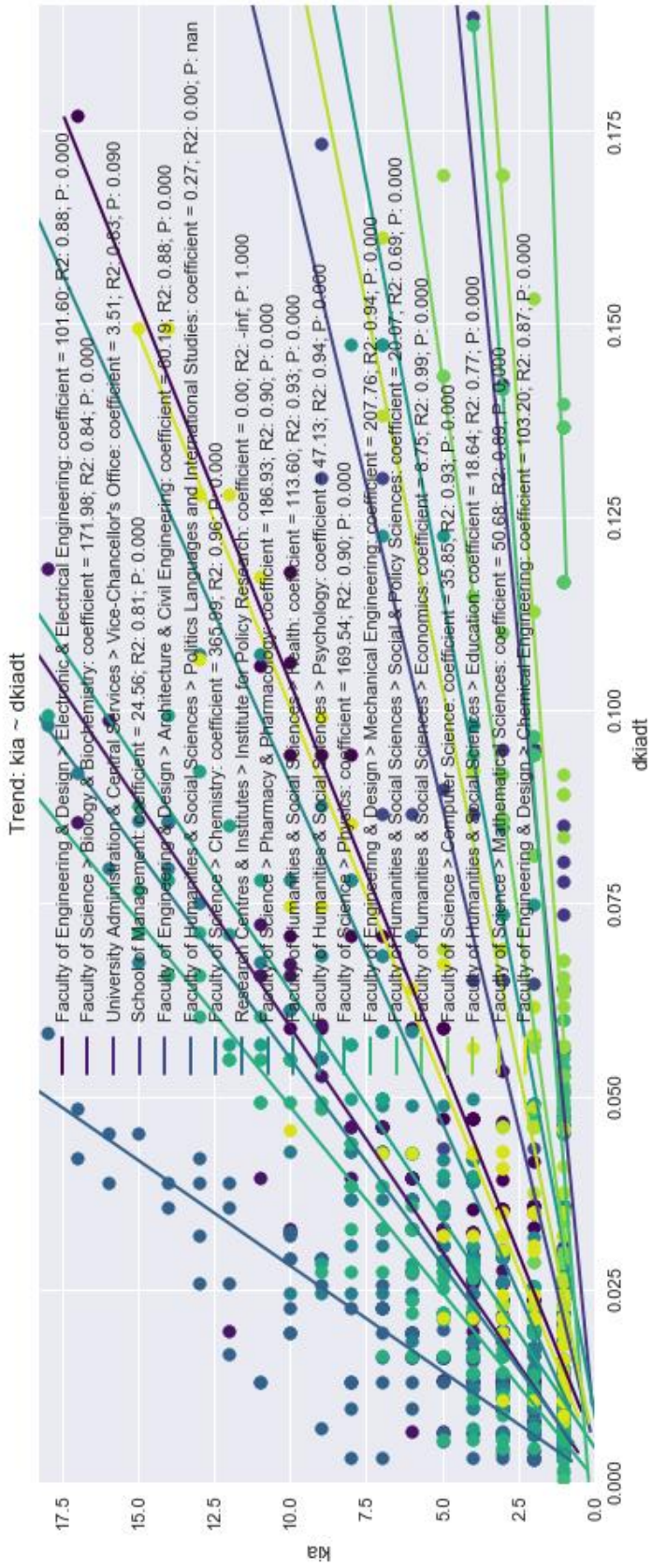

Figure 9.67. Correlation between the predictive model applied to the University of Bath 2000-2012 and used to predict connectivity of the University of Bath 2000-2013.

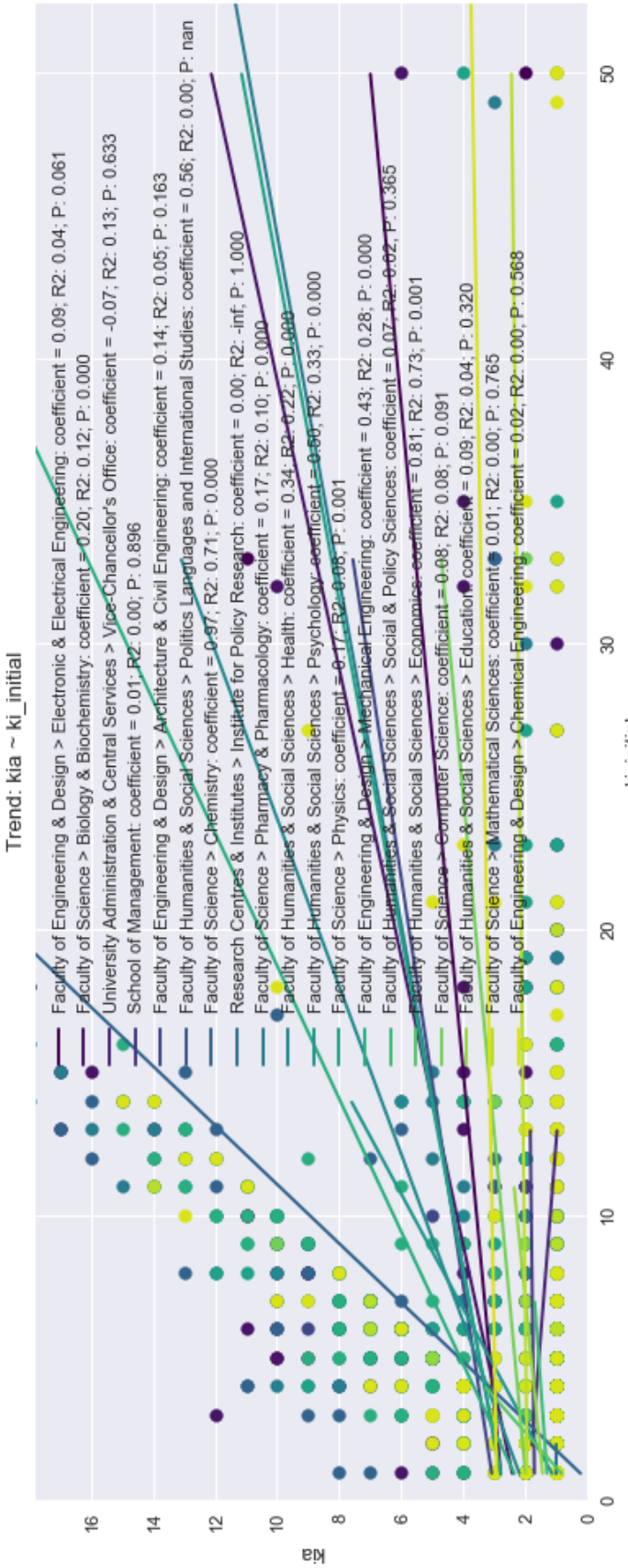

Figure 9.68. Correlation between the node degrees applied to the University of Bath 2000-2012 and used to predict connectivity of the University of Bath 2000-2013.

The plot for the R^2-values, shows that the R^2-values for the model are significantly higher than the degree.

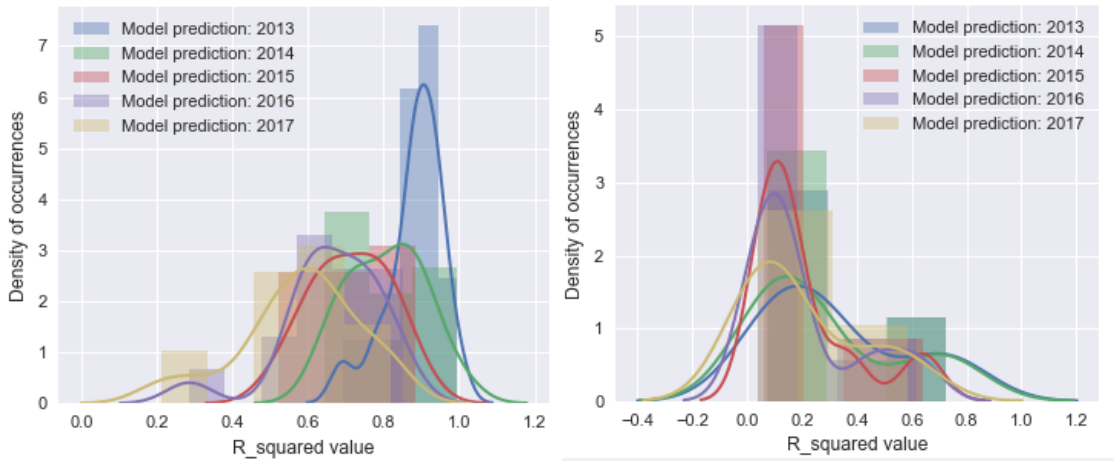

Figure 9.69. Distributions of the R^2-values correlating the multiplex model developed in this chapter (left) and the traditional network degree (right) to the future layer degrees for all node entities. The models are based on 2000-2012 values. The distributions show the histogram and the KDE plots.

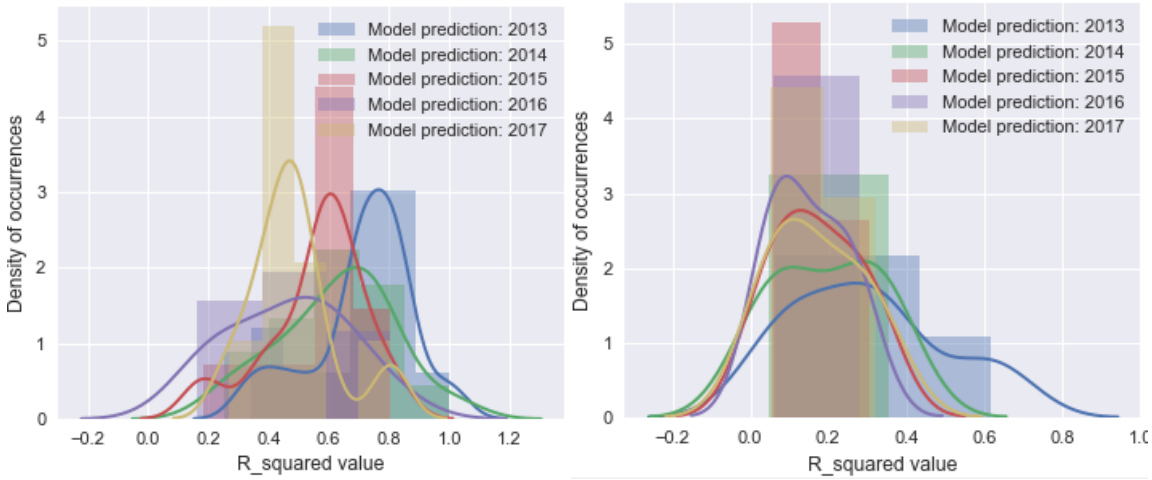

Figure 9.70. Distributions of the R^2-values correlating the multiplex model developed in this chapter (left) and the traditional network degree (right) to the future layer degrees for interdisciplinary node entities. The models are based on 2000-2012 values. The distributions show the histogram and the KDE plots.

Contribution to knowledge:

The predictive model created in this research is very capable statistically. For all node entities, the KDE peak R^2-values range from 0.91 to 0.60 from 1 to 5 years in the future as shown in Figure 9.69 (left). In comparison to degrees (a standard approach), which only predicts with KDE peak R^2-values ranging from 0.21 to 0.09 from 1 to 5 years in the future as shown in Figure 9.69 (right).

This model is therefore a very significant improvement on standard approaches, and provides the best predictive probability in comparison to review literature and analogous studies.

For IDR specifically, it is necessary to only look at interdisciplinary node entities only as shown in Figure 9.70. The KDE peak R^2-values range from 0.76 to 0.48 for the model, whereas the aggregate degree posts KDE peak R^2-values 0.27 to 0.09.

Hypothesis 9.6 is corroborated for up to four years in the future, and Hypothesis 9.7 is entirely corroborated. This therefore provides conclusive proof that the model developed is significantly superior at predicting future IDR in comparison to traditional networks models.

More importantly, this provides conclusive proof of the necessity of a multiplex framework, and that the output of the model serves as a good foundation for multiplex collaboration networks.

9.9. Chapter discussion

First and foremost, the predictive validation shows a massive improvement in predictive capability over standard approaches (which are not suitable) to identifying the future leaders of IDR who can enable and sustain it.

This has been the research aim, which can now be considered achieved.

However, whilst the research aim has been achieved, there is a lot of information that is embedded in the developed models and the results that require discussion and is pertinent to the research aim.

This chapter has developed a series of metrics to represent the overall structure of multiplex networks created using the framework developed in Chapter 8. These metrics defined the structure the University of Bath multiplex co-authorship network using department-based layers and content-based layers. These were determined to be highly similar.

Having developed these, it was possible to create a growth model to simulate the evolution of the network. The idea behind this was to uncover hidden mechanisms and incorporate them to the Barabási-Albert model in order to make it suitable for multiplex networks.

As such, a growth model was iteratively built up to be a good representation of the real University of Bath multiplex network. This allowed two things. It uncovered the hidden mechanics to provide insight into what might be causing the real-world network to look as it does. It also provided a mathematical representation that provides further insight into the dynamics of the network.

The multiplex framework developed was created as a way to define a node-centric multiplex collaboration network. The adopted framework necessitated node entities to exist in disciplines other than nodes' core disciplines. This turned out to be central to the model. It is worth noting that these could be represented in other frameworks, but occur naturally in the chosen framework.

Whilst there is a divide between disciplinary node entities and interdisciplinary node entities, every node entity provides a meaningful mechanism that is vital to defining the multiplex structure, as evidence by Models 3 and 4.

This brings us to the overall structure of the model: the degree in a layer is more important than the aggregate degree.

This implies that every person's presence in a different discipline can almost be considered a separate person. It is from this node entity that an individual builds further interdisciplinary research. This in itself is a highly valuable finding, although it is an indictment of human nature's difficulty in crossing thresholds to conduct IDR, both in department-based and content-based disciplines.

However, it became deficient in the layer-pair closeness, and created a Gaussian distribution instead of the power-law distribution with a negative exponent. This highlighted the importance of layer similarity (e.g. Chemistry and Physics will have more overlap than Chemistry and Sociology).

However, models based on layer-pair closeness grew Gaussian layer-pair closeness distributions. A conjecture was therefore formed that this closeness was not necessarily to do with the overlap in subject matter as much as it was an overlap in paradigm, and something that could be developed over time. Therefore, a layer's closeness centrality was determined to be a proxy for how easily a field was able to share, adopt, and adapt paradigms with other layers. This brings the focus into the second term of the developed model. This is that there is fundamentally a better suitability for IDR in one discipline compared to another. This turned out to be a vital mechanic.

This ultimately proved to be successful, although the reasoning remains conjecture.

All that can be said with utter certainty is that which has been corroborated in the hypotheses. The model provides a good simulation of the multiplex network structure (albeit not perfect), and that this model is a better predictor than the aggregate degree for IDR in the University of Bath department-based multiplex co-authorship network.

Despite the success, this work has several weaknesses as well.

The metrics chosen were intended to represent the overall structure. However, multiplex networks have focused entirely on developing analogous measures to traditional network science. Therefore, there is a clear divide between the 'horizontal' components, and the 'vertical' ones. There is undoubtedly information lost here, which could be vital to simulating the overall structure. This represents a very large gap in current knowledge.

Furthermore, the measures chosen were based on contemporary literature (Nicosia, Bianconi et al. 2013, Nicosia and Latora 2015), as well as established measures that significantly alter the structure of the network (Barabási and Pósfai 2016). However, other well-established measures were not included. For instance, clustering is noted to be absent. The reasoning behind this is that multiplex clustering is currently being established, and introducing such an element to the model would have made the implementation far more difficult. However, at the first opportunity, this research should be taken further forwards to develop a true understanding of how multiplex collaboration networks look structurally.

This research is also entirely networks based and only takes into consideration context based on what discipline individuals were classified in. This is a weakness of all networks work, and this research should be supported from the many different fields that overlap it. From Sociology, studies that could further our understanding of how it is that IDR is undertaken could complement this research or help outline its flaws. Sociological and Psychological studies could be performed to understand that the motivations of undertaking IDR. Policy Research could help us understand these motivations further. Soft Operations Research can help apply this work, in policy and strategy discussions between stakeholders. Ultimately, tackling IDR will require a large host of additional work to create a complete overview of how best to approach it.

9.10. Chapter Summary

This chapter sought to develop a model that could identify individuals who can enable and sustain IDR. It took the framework that was developed bespoke for collaboration networks in Chapter 8, and created a set of structural metrics with a view that if a network can be simulated, it can uncover hidden mechanics about how such a network is formed (Barabási and Pósfai 2016).

The model was built iteratively from the ground up to ensure that a simplistic model is created, and not one that is overfit.

The model was dependent on three terms: $\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}}$, B_i, and $\frac{{Q_{\alpha}}^{\gamma}}{\sum_{\mu}^{M}({Q_{\mu}}^{\gamma})}$. There was a large amount of dependency on the existence of node entities, which naturally drive node activity to a scale-free distribution.

This was validated in that it matched the historical data well. It was further validated and shown to have excellent predictive capabilities in who undertakes IDR in the future using a predictive validation approach.

Through this, the research aim was achieved.

Chapter 10: Research Discussion

This chapter reviews the main findings of the research. It does so with a view to discuss methods and assumptions, the overall validity of the methods chosen, and with the tempered view, discuss the overall implication of the findings.

Section 10.1 discusses the dataset collection methods. It reviews the validity of the findings in Chapters 7 and 9 with this in mind and discusses the sensitivity to the data fidelity. The section also discusses the definition of disciplines methods established in Chapter 6. It reviews the impact that erroneous classifications can have and discusses the importance of differing between department-based disciplines and content-based disciplines. Section 10.2 discusses the findings of Chapter 7, and what it ultimately means for the fields from which the models were drawn. Section 10.3 discusses the framework that was proposed and the validity of this approach over others, and whether this had a significant impact on the research outcome. Section 10.4 discusses the developed multiplex model, and the validity of its findings. Section 10.5 discusses the assumption that multiplex collaboration networks can be represented by multiplex co-authorship networks, which has strong implications for how extensible this research is. Section 10.6 outlines the implications of this research on the study of IDR, networks theory, and provides an outline for how this work can be used. Finally, section 10.7 reviews the discussion and makes concluding remarks.

10.1. The University of Bath Co-authorship dataset

The University of Bath co-authorship dataset was scraped using an open-source framework from the University of Bath opus. This was done in lieu of other data collection methods because the work needs to be extensible to other organisations and therefore requires a method that was applicable to other similar organisations.

As further network datasets are required to more extensively corroborate the findings in this thesis, having a method that can easily be applied in similar organisations is highly valuable.

However, the dataset relies entirely on the source being well-maintained, complete, and without bias. This unlikely to ever be the case, and sources of error stemming from the outset will ripple through the rest of the research.

As such, it is important to establish how valid was the method and outcome, and to establish how much effect it had on the overall research. This section establishes that due diligence was done in the method and validation, and any sources of error had minimal effect.

10.1.1. Method

The method of data collection took a standard scraping approach from the official University of Bath opus. It first established a way to identify unique IDs for every University of Bath author in order to ensure that similar names would not skew the results. It then scraped all the relevant publication data, which included the abstract text, department and centre information, and journal information.

The scraping framework itself ensured that no page was visited twice, meaning that unique information was pulled every time.

In terms of verification of the method and implementation, a full-stack developer with fluency in Python and scraping methods was consulted. The third-party verification greenlighted the implementation, confident that the method was working as intended.

The only issue with the method is that it collects author data from the publication data. This should strictly speaking be done separately (i.e. author data should be established from a source directly about the author if possible). The consequence is that effort was required to identify what discipline individuals belonged to.

This remains one of the biggest sources of error in the work.

10.1.2. Dataset validity

With regards to the dataset itself, the fact that scale-free distributions have been identified in a large number of different datasets, including many social and co-authorship networks (Newman 2010, Barabási and Pósfai 2016) provides the data validation that is required. No social or co-authorship networks with any other type of degree distributions have been found.

Therefore, the fact that the networks produce scale-free degree distributions provides a comparative validation as outlined in Chapter 5. Without prior knowledge of exactly what the network structure is, the fact that a scale-free distribution is the best data validation that can be achieved.

This validation assumes that the many previous examples of social networks holds true for organisation-based co-authorship networks. As this a very reasonable assumption, the dataset is deemed valid as a traditional network structure.

The difficulty arises when the multiplex network structure is analysed as the method to determine individuals' core-disciplines is inferred from the publication data as opposed to explicitly determined. There will be two types of nodes that are going to be less affected by this. Nodes with many publications, where the majority discipline classification will be used, and paradoxically nodes with only one publication (or few publications but all from the same discipline).

Whilst this method was validated for the department-based approach with 98.4% accuracy, the content-based approach was more difficult to determine. A good or better classification occurred only 69% of the time, although this accuracy increases every time multiple papers from the same individual are classified in the same discipline. Furthermore, as was stated, the entire purpose was to determine what discipline they were in, and therefore, if they fit well between two or more different disciplines, it is difficult to say they belong in one and not the other. However, content-based node classification does not benefit from the paradoxical benefit of single publication authors, and these remain the greatest source of inaccuracy, especially considering the scale-free nature of networks.

Therefore, the content-based classification is likely fraught with inaccuracy and needs greater validation or higher accuracy. It was for this reason that the research focused mostly upon department-based classification.

Greater accuracy and validation for content-based classification should be considered as future work.

10.2. The University of Bath discipline traditional networks models

Having discussed the dataset and its implications, it is possible to discuss the traditional network analysis in Chapter 7. The work in this chapter sought to adapt relevant SNA models to co-authorship networks with a specific focus on identifying the differences between disciplinary and interdisciplinary authors specifically.

Virtually none of the models were able to detect differences between disciplinary and interdisciplinary authors. The only model that was able to detect a statistically significant difference was using the content-based disciplines, which as discussed are highly inaccurate. Furthermore, the entire premise of disciplinary and interdisciplinary authors assumes that there are such archetypes. Finally, this study did not differentiate between contextual collaborations. Therefore, all correlations simply sought to find out whether they were successful or not, and could therefore not identify individuals who enable and sustain IDR.

These were the major flaws in the study.

10.2.1. Method

The method was based on a correlational study. It performed a panel analysis of the temporal data 2000-2010 to 2000-2017.

The overall analysis was a sound approach, but suffered from two major flaws.

The first flaw was the way that disciplinary and interdisciplinary authors were categorised (based on a threshold proportion of neighbours being interdisciplinary).

The second flaw is fatal and fundamental. The network framework adopted for the research could not make a distinction between different types of degrees. Therefore, any correlation would just compare its overall model to its overall metric of success. Even if authors could be considered interdisciplinary it would ultimately make no difference, as the prediction is not.

However, even finding this out has been immensely valuable. If the first step in moving in the right direction is realising you were wrong, the second step would be realising why you were wrong.

10.2.2. Implications of the study

The study highlighted the weaknesses of the traditional networks approaches to analysing multiple different disciplines. However, for the University of Bath, the study confirmed that the degree, the betweenness, and the PageRank centralities all provided a positive correlation with academic output. The structural holes measure also provided a positive correlation to academic output. The strength of weak ties was rejected as a model for correlating to academic output.

These are all important findings within their own right, as they provide evidence for the applicability of these models in research organisation-based co-authorship networks. Future researchers can use these findings in direct or analogous research.

The degree centrality model was corroborated and agreed with several different studies (McFadyen and Cannella 2004, McFadyen and Cannella 2005, McFadyen, Semadeni et al. 2009). The betweenness centrality model was also corroborated with previous studies (Li, Liao et al. 2013). The eigenvector (PageRank) centrality found opposing trends to what has been reported in some literature, although it is important to realise that the literature was mixed on this model (Cimenler, Reeves et al. 2014). The structural holes model was improved upon to include second order structural holes (open rectangles) and found that there was good agreement with theory (Burt 2004, Burt 2009). Finally, the strength of weak ties has been disproved for the co-authorship networks (Granovetter 1973).

The study also disproves the disciplinary and interdisciplinary archetypes. As no differences could be found there is no evidence that these archetypes exist, thereby necessitating an improved network framework to study IDR.

10.2.3. Further work

The correlational study was broad and yielded a lot of interesting results. It would be necessary to adapt the models to their multilayer counterpart and conduct a similar study again.

10.3. The multiplex collaboration network framework

Having identified the difficulty of modelling IDR using traditional networks, the next steps in the research required an approach that could include multiple different types of links. The multiplex collaboration framework was developed to establish how it is that such a multilayer network could be created using the dataset. Two major approaches were established in the literature: multiplex and network-of-network frameworks. The network-of-networks framework was rejected due to it requiring a measurable difference between disciplinary and interdisciplinary links, thereby adding additional assumptions to the research.

The multiplex network is also a better representation of the multiple different links that occur and is therefore more studied. However, the decision to make the multiplex implementation a multilayer instead of multi-edge framework is perhaps the most central decision to the entire thesis.

The multi-edge framework would simply assume a traditional network but differ on the types of links. This could easily be accommodated by the fact that it is the links that are classified, thereby removing the assumptions on how it is that nodes are classified.

However, it does not provide any structure to what an entire discipline is (e.g. a department). The choice to adopt a multilayer, node-centric classification was done on the basis of trying to emulate an organisational structure. The existence of node entities becoming at first a mathematical curiosity, and subsequently a central element to the final model.

The discovery that node entities could drive the dynamics of a multiplex network structure was formalised with growth model with randomly assigned links between nodes being able to successfully recreate a scale-free node activity distribution provided a significant original contribution to knowledge.

Therefore, the trade-off of choosing to keep the node classified paradigm, to enable a node entity framework over the multi-edge paradigm, is deemed to be worth it.

The node centric framework of course required a core-discipline, which has not been done before. This ultimately created two different types of node entities: disciplinary node entities and interdisciplinary node entities which allows for the differences between disciplinary research and IDR to be analysed.

10.4. The University of Bath multiplex co-authorship network

Having established a framework, it was possible to analyse the dataset and to develop a greater understanding of multiplex collaboration networks. Many different valid approaches to this are

possible. However, the philosophy adopted in this research was to try to establish a fundamental understanding as to why the networks evolve to look the way they do, and then codify this into a model that explicitly identify individuals who enable and sustain IDR. If a real network structure can be recreated through a growth model, then that growth model is able to uncover hidden mechanics about how it is that networks form.

The growth model needed to be validated, and two different validation paradigms were determined to be suitable. Historical validation served as a goal to be achieved. If the ultimate structure of the network were to match the historic structure, this would be an indication that some of the hidden mechanics on how multiplex networks form have been uncovered. To do so, this part of the research again took advantage of the deductive philosophy and hypothesis testing as a way of determining how similar two structures were.

A growth model was created that found good agreement with the historical values, and the following model was extracted from the growth model.

$$\frac{dk_i^{\alpha=D_i}}{dt} = A_0(B_i + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1A_2(B_i - 1)\,\langle N_t^{\beta}\rangle^{\beta\neq D_i}\,\langle\frac{1}{N_t^{\beta}}\frac{{Q_\beta}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)}\rangle^{\beta\neq D_i} \tag{10.1}$$

$$\frac{dk_i^{\alpha\neq D_i}}{dt} = A_0(1 + Mq)\frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} + A_1\frac{N_t^{D_i}}{N_t^{\alpha}}\frac{k_i^{D_i}}{\sum_{j=1}^{N_t^{D_i}} k_j^{D_i}}\frac{{Q_\alpha}^{\gamma}}{\sum_{\mu}^{M}\left({Q_\mu}^{\gamma}\right)} \tag{10.2}$$

This model was further validated by being able to provide excellent predictive capability for future IDR using the University of Bath department-based multiplex co-authorship network longitudinal data for its predictive validation. It achieved an R^2-value in the excess of 0.9 for predictions one year in the future and an R^2-value of 0.6 for five years in the future. No models have achieved such predictive capability before.

However, this prediction was purely based on how many future collaborators there are. It is necessary to establish how relevant this metric is and whether it is robust enough to engage stakeholders and decision makers in policy discussions.

It is of course important to understand that this is just a model that seeks to recreate behaviour observed in the real-world, and that through we hope to understand phenomena. That is to say, that this model's dynamics can only describe the behaviour the system as a whole (what it is validated for); it cannot be used to draw inference on individuals' motives, only on their opportunities.

In that respect, this model is subject to the concept of Falsificationism – the model will hold until it is refuted. It is for this reason that the most important future work should endeavour towards

falsifying this model. In the best-case scenario, our understanding of multiplex collaboration networks will be deepened, in the worst-case scenario, the model will hold.

10.4.1. Method and dataset validity

The method of gaining deeper insight into networks by providing a realistic simulation is arguably one of the most central methods of understanding network dynamics as the Barabási-Albert model heralded a resurgence of interest in network science (Barabási and Albert 1999, Newman 2010). The pioneering work alongside important discoveries of network properties such as the small-world phenomenon (Watts and Strogatz 1998) led to an upsurge in networks interest.

The reason the growth model method was chosen was based on trying to recreate such success on a fundamental level for multiplex collaboration networks. Barabási states in his book (Barabási and Pósfai 2016) that the success of the model was in uncovering hidden assumptions in the Erdős-Renyi model. In a similar way that traditional network growth models are not suited to recreate multiplex networks, this work hoped to uncover the hidden assumptions pertaining to multiplexity.

The question arises on what the uncovered mechanics represent. It is impossible to say the mechanics are causal. They may even be just confounding. However, if we accept this and understand that effort is required to uncover why these mechanics provide accurate multiplex structures, the model will remain useful.

These mechanics are not absolute, but do recreate the real-world multiplex structure. Anecdotally, by visualising the networks (www.hultin.uk/visualise), these mechanics do hold. Particularly the interdisciplinary mechanics, as interdisciplinary links exist between any individuals, and does not seem to favour highly connected nodes. This is particularly true for the researcher's immediate surroundings, which can be thought of as a face validity.

In the final analysis of the method, the work does not make any claims beyond what it is meant to do, or beyond what it is 'validated' for. That is to say, the model holds good agreement with historical structures, and has excellent predictive capabilities for IDR for the University of Bath co-authorship 2012-2017.

In order to extend the validity, the model needs to be corroborated by different datasets.

However, as it has stood the deductive tests outlined in this research, the model holds, and the research aim has therefore been achieved: an explicit predictive model that can identify the individuals who enable and sustain IDR.

10.5. The multiplex collaboration network model

This section discusses whether it is reasonable to assume that the co-authorship network is a good proxy for collaboration networks, and whether the findings for the co-authorship networks can be extended to all collaboration networks.

Popperian philosophy suggests that this is not possible, as it makes inferences and cannot therefore be considered scientific knowledge.

However, co-authorship was chosen to be the operational definition of collaboration in this research and is therefore just viewed as a measure of collaboration. It is unlikely that any multiplex collaboration network constructed using the same framework will differ significantly. However, this statement in itself is just a conjecture.

The view taken in this research is that the work is representative of collaborations as it has been corroborated by one operational definition. However, it is only weakly corroborated and requires further corroboration to increase the confidence that the theory holds.

10.6. Implications for IDR and networks theory

There is a very broad body of literature surrounding IDR, and growing interest regarding it academically, administratively, managerially, and strategically (Siedlok and Hibbert 2014). However, the studies surrounding IDR have been lacking in broad quantitative approaches. With big data becoming more readily available, cheap, objective, correlational, and longitudinal datasets are available to conduct various analyses upon. SNA was determined to be a suitable approach as it has been applied in analogous research (Wasserman and Faust 1994). However, it becomes apparent that traditional SNA approaches do not provide the resolution needed to distinguish IDR, and it was necessary to define disciplinary and interdisciplinary archetypes. These archetypes do provide statistically significant differences and were therefore rejected as a theory.

This provides an explanation why there is little research surrounding IDR using SNA methods. A multilayer perspective provides the needed analytical resolution, but is a fairly new field and is still in its early stages of defining its frameworks and statistical mechanics (Kivelä, Arenas et al. 2014). Never the less, this research proposed a framework tailored specifically for collaboration networks with different disciplines.

The framework has been successful and appropriate steps have been taken to ensure fidelity. Through this framework, a model was created that proposes how multiplex collaboration networks form and evolve, and through that individuals who can sustain and enable IDR are now identifiable with good predictive capability.

This is highly impactful research that addresses the needs outlined above. It provides a robust dataset that can be analysed quantitatively. This can allow stakeholders to make evidence-based decisions, or engage them in an objective representation of the research organisation's collaborative landscape. Such a conceptual model could serve as valuable discussion platform (Checkland 1999). The model itself provides evidence on what type of individuals most readily overcome the barriers to IDR, and highlights that this not a person trait (i.e. not an interdisciplinary person), but rather is due to the process (see section 9.3.1), due to their exposure (see section 9.6.8), and the readiness of their knowledge to be shared and adapted (see section 9.7). This means that finding individuals who can overcome each of these, are the individuals most likely to sustain and enable IDR.

It also provides applied networks theory with valuable information regarding the applicability of the various models on a research organisation with hard boundaries. There was a positive correlation to the academic output with the degree, betweenness, and eigenvectors centralities. There was also a positive correlation with the structural holes, and it was shown that the structural holes can be extended to include 2nd-order holes to provide a better correlation. Finally, evidence was shown to reject the strength of weak ties concept. This research also established that traditional network approaches are unsuitable to investigate differences in teams or disciplines, and a multilayer perspective is needed.

This research then laid down the foundations for analysing multiplex collaboration networks. The framework developed was shown to be highly dependent on the node entities, thereby showing that people still form strong barriers based on either department or content boundaries. This manifests itself in that node entities can be treated as semi-individual people in their own right (see section 9.6.8). It goes to show that node entities naturally create a scale-free node activity distribution, which can be considered the 'vertical' degree in multiplex networks. Finally, it identifies the layer closeness centrality as being vital to creating realistic multiplex structures, suggesting that some disciplines are better at IDR than others (e.g. Mathematics can be applied in almost all fields).

All of these findings advance the knowledge in these respective fields. However, it is important to note that the work is foundational, and needs further corroborating evidence.

10.7. Research aim

The overall research design has endeavoured to remain as simple as possible as the subject matter is complex. This is possibly a weakness as it may be oversimplifying the subject of IDR, which may need a more nuanced explanation. However, it has provided a strong predictive model, which was truly the aim of the research from the outset.

In that respect, the research design has been successful, and the research aim and research objectives have been achieved.

Chapter 11: Research Conclusion and proposed further work

This chapter seeks to conclude the overall work. It does so by summarising the contributions of the various chapters. It then reflects on how the research aim and objectives were achieved. This chapter then outlines the major contributions this research provides to advance our understanding of both IDR and networks theory. Finally, it outlines the further work required.

11.1. Chapter summaries

Chapter 1 introduced the research and discusses its context. The chapter outlined the major challenges and opportunities that IDR faces. It establishes that it is necessary to enable IDR to overcome these barriers.

Chapter 2 outlined the research method that guided the research. It defined a structure research methodology that guided the research. Part of this research methodology was the research onion, which helped design the research. The research was designed to represent the complex ideas, findings, and models in a structured manner. As such, a deductive research approach ensured that every finding in this research was testable. This may lose out on more qualitative findings, but neither the research aim, nor the dataset were amenable to inductive approaches.

Chapter 3 outlined the need to conduct quantitative research on how it is that IDR can be enabled and sustained. It outlined a broad literature review of the studies into IDR have found that most studies focused mainly on the barriers to IDR (Campbell 2005, Davidson 2015), discussions on methods and process (Repko 2008), and qualitative studies (Roche and Rickard 2017). Most of the quantitative studies focused on measuring IDR (e.g. H-index) (Yegros-Yegros, Rafols et al. 2015, Huutoniemi and Rafols 2016); there was a distinct lack of research in quantitative studies that focused on uncovering correlations and IDR mechanics through quantitative methods. As such, this research attempts to address this gap in knowledge, and chooses a SNA framework to do so.

Chapter 4 outlined major networks science measures that are needed to understand networks studies (Albert and Barabási 2002, Newman 2010, Barabási and Pósfai 2016). Having outlined these, it then discusses the broad and fragmented literature that exists surrounding IDR analogous studies. Some studies were specific to IDR (Yegros-Yegros, Rafols et al. 2015), but none were found that achieve the research aim. Analogous studies provided several models (Burt 2004, McFadyen and Cannella 2004, White, Wellman et al. 2004, McFadyen and Cannella 2005, Burt 2009, McFadyen, Semadeni et al. 2009)

Chapter 5 outlined the dataset requirements for this research, where appropriate data could be found, and how it was collected. It outlined the creation of scarping tool that was made bespoke for

collecting data suitable to construct a journal co-authorship network that would enable both cross-sectional and longitudinal analyses to be performed.

Chapter 6 outlined the difficulty of choosing an operational definition for disciplines. A department-based discipline definition was straight-forward to achieve. A content-based definition was also achieved, but with significantly worse accuracy that may have affected its results. Nevertheless, similar structural properties were identified for both definitions throughout the research. It is also important to realise that the content-based classification allows this work to be extensible to datasets that do not contain any department-based classifications. It is also useful to identify differences between the department-based classifications and content-based classifications, which should be conducted as further work.

Chapter 7 tests five models established in Chapter 4 using the dataset established in Chapters 5 and 6. The models were adapted to test disciplinary and interdisciplinary author archetypes. The study showed that there were no statistically significant differences, and that the archetypes were therefore rejected. This may explain why there are few IDR networks studies. However, the study provided valuable results: it found a positive correlation to authors' impact factor for the degree, betweenness, PageRank, and structural holes models. The strength of weak ties was rejected, as an opposite trend was found. Ultimately, this study suffered two major flaws: it assumed that there were archetypes, and it correlated to an overall measure. To create findings specific to IDR it was found to be necessary to extend the network framework to a multilayer perspective.

Chapter 8 reviews multilayer networks. As it is a nascent field, the research has not yet reached the rigour, structure, and unity that traditional network science does. For instance, there are many different ways to formulate the same structure. The review showed that multilayer structures significantly alter the dynamics of a network, and are therefore important to take into consideration (Gomez, Diaz-Guilera et al. 2013). As it is a nascent field, and few studies could be found regarding multilayer collaboration networks, a bespoke framework was defined for this research, and a decision was made to attempt to recreate the success in understanding network structures through the use of growth models for this multiplex framework.

Chapter 9 establishes several metrics that describe multiplex structures. Exemplar structures were established from the University of Bath multiplex co-authorship networks 2000-2017. These provided the historical data validation that the growth models sought to structurally recreate. By iteratively building up the model, several interesting and impactful findings were uncovered that provided original contributions to knowledge. The greatest of these were:

- The identification that as node entities behave differently, but belong to the same person thereby implying that any differences that were found occurred due to a difference in process, not the person.
- The importance of node entities in IDR in forming power-law distributions for the node activity, which has been shown to act as a 'vertical' degree measure.
- The importance of layer closeness centrality, without which the growth models could not recreate a realistic multiplex structure. Implying that certain disciplines are simply better at sharing and adapting knowledge to/from other disciplines and conducting IDR.
- A mathematical expression for the rate at which individuals will develop interdisciplinary collaborators: an operational definition for enabling and sustaining IDR. Thereby achieving the research aim.

Chapter 10 discusses the overall thesis and tempers the findings by critically reviewing them. It then discusses the impact that these findings have.

11.2. Research aim and objectives

The research aim was defined as follows.

> *To create a model that identifies individuals who enable and sustain interdisciplinary research.*

Based on the predictive validation conducted in Chapter 9, this was achieved. Good predictive capabilities for IDR up to four years in the future were obtained.

Objective 1. To choose an appropriate and useful approach to analysing IDR from a people centric perspective.

SNA was the approach taken initially, but as was shown in Chapter 7. This approach was unsuitable and required the networks framework to be extended.

Objective 2. To define and collect a dataset that suits the needs of the chosen approach.

This objective was achieved in Chapters 5 and 6. The requirements of the dataset were considered, and the data was validated.

Objective 3. To establish the validity of prevailing models in IDR literature and analogous research in the collected dataset.

Chapter 3 and 4 established approaches to IDR and prevailing models in analogous research respectively.

Objective 4. To develop a framework that addresses the deficiencies of the prevailing models in IDR literature and analogous research.

Chapter 7 outlined the deficiencies of the prevailing models, and multilayer frameworks were reviewed in Chapter 8. Chapter 8 then proposed a framework that suited the needs of the research.

Objective 5. To develop a model using the framework to achieve better predictive capability in identifying the future leaders of IDR in comparison to standard approaches.

Chapter 9 developed a series of growth models. Model 4 achieved good agreement with exemplar multiplex networks. Model 4 was then used to extract a mathematical representation for identifying the future leaders of IDR.

Objective 6. To validate the model using the collected dataset.

Chapter 9 validated the growth models through historical data agreement, and validated the mathematical expression through predictive validation. As both models are interlinked, this can be thought of as two forms of validation.

Objective 7. To discuss the strengths, weaknesses, and implications of the created model.

The strengths, weaknesses, and implications of the created model were discussed in sections 9.7.8, 9.8, 9.9, and Chapter 10 in varying levels of aggregation.

11.3. Research contributions

This research has trodden new ground in both application and theory. Through the thesis, this research has attempted to draw attention to significant research contributions that have implications on application or theory, or provide re-useable data and information.

The following significant contributions to knowledge have been identified.

1. This research outlined the needs of a dataset to perform multilayer collaboration network analyses upon. It also created a tool that can collect such data from similar sources. This is very important as it allows a series of datasets to be created and for the research to be further validated. The contribution to knowledge is both in terms of the process (defining what the dataset is) and in actually providing the dataset that can easily be shared with peers. This data has been validated (see section 5.6).
2. The research established that the various models from analogous research held, but that these were not adaptable to IDR using a traditional networks framework. The disciplinary and interdisciplinary archetypes were refuted as a hypothesis.
3. The unsuitability of traditional networks approaches to investigate IDR and other networks studies that may classify nodes is an important contribution to knowledge, as it shows the importance of adopting a multilayer perspective.
4. The research proposed a multiplex framework as a rank-3 tensor of format $N \times N \times M$. This was done in lieu of the rank-4 tensor of format $N \times N \times M \times M$. This ensured that disciplines could be compared as otherwise there would be $M \times M$ layers, and no overlap. To establish that there is a link from one discipline to another, and compare its local structure (e.g. the equivalent of layer degree) two separate operations would have to be performed. Furthermore, the rank-3 tensor format created node entities in only M layers, thereby highlighting their effect as individual entities within disciplines.
5. The difference between disciplinary and interdisciplinary node entities were found across all results. As node entities represent parts of a node (an author), it becomes clear that interdisciplinary structures occur due to the process, not the person.
6. A trend reversal in degree-correlation from the multiplex network to the aggregated network was found. This behaviour had not been previously identified within published literature. This trend reversal has been identified in unpublished work (Hultin, unpublished) since the main body of work in this thesis using the same growth model process including an interdisciplinary link removal (not included in this thesis). This thesis however merely identifies that such a trend occurs.
    a. A different analytical analysis using a forward time stepping scheme to calculate the degree-correlation distribution compared to what has been reported in literature provides a better approximation of the Barabási-Albert model. This finding is however tangential.
7. The node activity can be thought of as a ‘vertical’ node degree and forms a negative power-law distribution. The node entities were found to be vital to the process and naturally cause the node activity to tend towards realistic structures. This implies that node entities form a vital component of multiplex networks. This has far-reaching

consequences as the node entities can be treated as quasi-individuals. This suggests that the barriers between the layers are very steep as no little benefit can affect a node entity from outside the discipline.

8. The layer closeness centrality forms a vital part of creating a realistic network structure. This suggests that some disciplines are inherently better at sharing their knowledge than others. This could be due to the subject matter (e.g. Mathematics) being directly applicable in IDR, or that the research paradigms in the discipline are easily adapted and shared to better conduct IDR.
9. A mathematical formulation for the rate of change of layer degree provided a model that can directly be used to identify individual who enable and sustain IDR. This was validated to have good predictive capabilities for IDR, and found a good correlation (R^2-value > 0.5) for up to four years.

11.4. Further work

Multilayer network measures have thus far created measures that investigate the 'horizontal' (analogous to traditional network measures) and 'vertical' (node-aligned measures such as node activity) measures, but few measures exist that study the combination of these that may be required to truly appreciate the structure of multiplex networks. It is necessary to significantly extend the tools available to analyse such three-dimensional structures. This affects this research directly as the growth model was based on available measures, and may therefore not be a suitable overall representation of the University of Bath multiplex co-authorship network structure.

The work has also thus far only been corroborated for one dataset. In order to strengthen its validation, this work should be conducted for other organisations. Data has already been collected for the Cranfield University, which is ready to be analysed. This represents the immediate next steps in the research.

The research also requires a practical implementation. A model without individuals using it remains just an artefact. The model therefore needs to be engaged with. Therein lies networks' greatest strength: the ability to visualise a system. By virtue of multiplex networks adding another dimension (through its layers), the visualisation and stakeholder engagement needs to be reviewed.

It would be useful to unify some of the models established in analogous works in traditional SNA with a multilayer perspective. The developed framework and model represents a powerful tool that can be used to analyse a variety of different phenomena.

Finally, the aim of identifying individuals who enable and sustain IDR was achieved. However, the research outlined here has described how it is that we can detect it and has discussed the implications

of the various findings on how it is affected by the IDR process. However, to achieve fuller understanding of IDR, many different perspectives are needed to establish motives, culture, policy, epistemic factors, and likely countless more.

It becomes clear that the push for IDR is not just a necessity; it is a path that leads us to break down epistemic barriers that we have set ourselves.

> *"These rules, the sign language and grammar of the Game, constitute a kind of highly developed secret language drawing upon several sciences and arts, but especially mathematics and music (and/or musicology), and capable of expressing and establishing interrelationships between the content and conclusions of nearly all scholarly disciplines. The Glass Bead Game is thus a mode of playing with the total contents and values of our culture; it plays with them as, say, in the great age of the arts a painter might have played with the colours on his palette."*
>
> -Herman Hesse, The Glass Bead Game

Bibliography

Abbasi, A., J. Altmann and L. Hossain (2011). "Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures." Journal of Informetrics **5**(4): 594-607.

Abbasi, A., K. S. K. Chung and L. Hossain (2012). "Egocentric analysis of co-authorship network structure, position and performance." Information Processing & Management **48**(4): 671-679.

Abrahamsson, P., O. Salo, J. Ronkainen and J. Warsta (2017). "Agile software development methods: Review and analysis." arXiv preprint arXiv:1709.08439.

Ackoff, R. L. (1979). "Resurrecting the future of operational research." Journal of the Operational Research Society: 189-199.

Aggarwal, C. C. and C. Zhai (2012). A survey of text clustering algorithms. Mining text data, Springer**:** 77-128.

Albert, R. and A.-L. Barabási (2002). "Statistical mechanics of complex networks." Reviews of Modern Physics **74**(1): 47-97.

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut (2017). "Text summarization techniques: A brief survey." arXiv preprint arXiv:1707.02268.

Amaral, L. A. N., A. Scala, M. Barthelemy and H. E. Stanley (2000). "Classes of small-world networks." Proceedings of the national academy of sciences **97**(21): 11149-11152.

Anauati, V., S. Galiani and R. H. Gálvez (2016). "Quantifying the life cycle of scholarly articles across fields of economic research." Economic Inquiry **54**(2): 1339-1355.

Ancona, D. G. and D. F. Caldwell (1992). "Bridging the boundary: External activity and performance in organizational teams." Administrative science quarterly: 634-665.

Andersen, D. F., G. P. Richardson and J. A. Vennix (1997). "Group model building: adding more science to the craft." System Dynamics Review **13**(2): 187-201.

Andersen, H. and B. Hepburn (2016). "Scientific Method." The Stanford Encyclopedia of Philosophy.

Arthur, D. and S. Vassilvitskii (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics.

Arthur, W., J. Holland, B. LeBaron, R. Palmer and P. Tayler (1996). "Asset Pricing Under Inductive Reasoning in Artificial Stock Market." SFI paper in progress.

Atkinson, R. D. and S. J. Ezell (2012). Innovation economics: the race for global advantage, Yale University Press.

Audia, P. G. and J. A. Goncalo (2007). "Past success and creativity over time: A study of inventors in the hard disk drive industry." Management Science **53**(1): 1-15.

Backmann, J., M. Hoegl and J. L. Cordery (2015). "Soaking it up: Absorptive capacity in interorganizational new product development teams." Journal of Product Innovation Management **32**(6): 861-877.

Bacon, F. (1864). "Meditationes Sacrae (1597)." The Works of Francis Bacon **14**.

Baltagi, B. (2008). Econometric analysis of panel data, John Wiley & Sons.

Banerjee, A., A. G. Chandrasekhar, E. Duflo and M. O. Jackson (2014). Gossip: Identifying central individuals in a social network, National Bureau of Economic Research.

Barabasi, A.-L. (2005). "The origin of bursts and heavy tails in human dynamics." Nature **435**(7039): 207-211.

Barabási, A.-L. and R. Albert (1999). "Emergence of scaling in random networks." science **286**(5439): 509-512.

Barabâsi, A.-L., H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek (2002). "Evolution of the social network of scientific collaborations." Physica A: Statistical mechanics and its applications **311**(3-4): 590-614.

Barabási, A.-L. and M. Pósfai (2016). Network science, Cambridge university press.

Barrat, A., M. Barthelemy, R. Pastor-Satorras and A. Vespignani (2004). "The architecture of complex weighted networks." Proceedings of the National Academy of Sciences of the United States of America **101**(11): 3747-3752.

Barrat, A., M. Barthélemy and A. Vespignani (2004). "Modeling the evolution of weighted networks." Physical review E **70**(6): 066149.

Barrat, A. and R. Pastor-Satorras (2005). "Rate equation approach for correlations in growing network models." Physical Review E **71**(3): 036127.

Barrios, F., F. López, L. Argerich and R. Wachenchauzer (2016). "Variations of the similarity function of textrank for automated summarization." arXiv preprint arXiv:1602.03606.

Barry, A., G. Born and G. Weszkalnys (2008). "Logics of interdisciplinarity." Economy and Society **37**(1): 20-49.

Barthélemy, M., A. Barrat, R. Pastor-Satorras and A. Vespignani (2004). "Velocity and hierarchical spread of epidemic outbreaks in scale-free networks." Physical Review Letters **92**(17): 178701.

Barthélemy, M., B. Gondran and E. Guichard (2002). "Large scale cross-correlations in Internet traffic." Physical Review E **66**(5): 056110.

Bartneck, C. and S. Kokkelmans (2011). "Detecting h-index manipulation through self-citation analysis." Scientometrics **87**(1): 85-98.

Bartunek, J. M. (2007). "Academic-practitioner collaboration need not require joint or relevant research: Toward a relational scholarship of integration." Academy of Management Journal **50**(6): 1323-1333.

Barzel, B. and A.-L. Barabasi (2013). "Universality in network dynamics." Nat Phys **9**(10): 673-681.

Batallas, D. A. and A. A. Yassine (2006). "Information leaders in product development organizational networks: Social network analysis of the design structure matrix." IEEE Transactions on Engineering Management **53**(4): 570-582.

Battiston, F., V. Nicosia and V. Latora (2014). "Structural measures for multiplex networks." Physical Review E **89**(3): 032804.

Bavelas, A. (1950). "Communication patterns in task-oriented groups." Journal of the acoustical society of America.

Becker, G. S. (2013). The economic approach to human behavior, University of Chicago press.

Benckendorff, P. and A. Zehrer (2013). "A network analysis of tourism research." Annals of Tourism Research **43**: 121-149.

Berlingerio, M., M. Coscia, F. Giannotti, A. Monreale and D. Pedreschi (2011). "The pursuit of hubbiness: analysis of hubs in large multidimensional networks." Journal of Computational Science **2**(3): 223-237.

Berlingerio, M., M. Coscia, F. Giannotti, A. Monreale and D. Pedreschi (2013). "Multidimensional networks: foundations of structural analysis." World Wide Web **16**(5-6): 567-593.

Bianconi, G. and A.-L. Barabási (2001). "Competition and multiscaling in evolving networks." EPL (Europhysics Letters) **54**(4): 436.

Bird, S., E. Klein and E. Loper (2009). Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.".

Bird, S., E. Loper and E. Klein "Natural Language Toolkit." from https://www.nltk.org/.

Blessing, L. T. and A. Chakrabarti (2009). DRM: A Design Reseach Methodology, Springer.

Blockley, D. I. and P. Godfrey (2000). Doing it differently: Systems for rethinking construction, Thomas Telford.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre (2008). "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment **2008**(10): P10008.

Bollobás, B. (1981). "Degree sequences of random graphs." Discrete Mathematics **33**(1): 1-19.

Bompard, E., B. Han, M. Masera and E. Pons (2014). Smart grid as multi-layer interacting system for complex decision makings. Networks of Networks: The Last Frontier of Complexity, Springer**:** 187-201.

Bonacina, F., M. D'Errico, E. Moretto, S. Stefani, A. Torriero and G. Zambruno (2015). "A multiple network approach to corporate governance." Quality & Quantity **49**(4): 1585-1595.

Bordons, M., J. Aparicio, B. González-Albo and A. A. Díaz-Faes (2015). "The relationship between the research performance of scientists and their position in co-authorship networks in three fields." Journal of Informetrics **9**(1): 135-144.

Bornmann, L. and H. D. Daniel (2008). "What do citation counts measure? A review of studies on citing behavior." Journal of Documentation **64**(1): 45-80.

Bothner, M. S. (2003). "Competition and social influence: The diffusion of the sixth-generation processor in the global computer industry." American Journal of Sociology **108**(6): 1175-1210.

Bozeman, B. and E. Corley (2004). "Scientists' collaboration strategies: implications for scientific and technical human capital." Research Policy **33**(4): 599-616.

Brandes, U. (2001). "A faster algorithm for betweenness centrality." The Journal of Mathematical Sociology **25**(2): 163-177.

Breiger, R. L. and P. E. Pattison (1986). "Cumulated social roles: The duality of persons and their algebras." Social networks **8**(3): 215-256.

Bródka, P., P. Kazienko, K. Musiał and K. Skibicki (2012). "Analysis of neighbourhoods in multi-layered dynamic social networks." International Journal of Computational Intelligence Systems **5**(3): 582-596.

Bródka, P., K. Skibicki, P. Kazienko and K. Musiał (2011). A degree centrality in multi-layered social network. Computational Aspects of Social Networks (CASoN), 2011 International Conference on, IEEE.

Bruce, A., C. Lyall, J. Tait and R. Williams (2004). "Interdisciplinary integration in Europe: The case of the Fifth Framework programme." Futures **36**(4): 457-470.

Bryman, A. and E. Bell (2015). Business research methods, Oxford University Press, USA.

Buck, S. (2015). Introductory Applied Econometrics, EEP/IAS 118.

Buldyrev, S. V., R. Parshani, G. Paul, H. E. Stanley and S. Havlin (2010). "Catastrophic cascade of failures in interdependent networks." Nature **464**(7291): 1025-1028.

Bunderson, J. S. and K. M. Sutcliffe (2002). "Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects." Academy of management journal **45**(5): 875-893.

Burt, R. S. (2004). "Structural holes and good ideas." American journal of sociology **110**(2): 349-399.

Burt, R. S. (2009). Structural holes: The social structure of competition, Harvard university press.

Butler, S. K. (2008). Eigenvalues and structures of graphs, ProQuest.

Cai, D., Z. Shao, X. He, X. Yan and J. Han (2005). Community mining from multi-relational networks. European Conference on Principles of Data Mining and Knowledge Discovery, Springer.

Campbell, L. M. (2005). "Overcoming obstacles to interdisciplinary research." Conservation biology **19**(2): 574-577.

Carayol, N. and T. U. N. Thi (2005). "Why do academic scientists engage in interdisciplinary research?" Research evaluation **14**(1): 70-79.

Cardillo, A., J. Gómez-Gardenes, M. Zanin, M. Romance, D. Papo, F. Del Pozo and S. Boccaletti (2013). "Emergence of network features from multiplexity." Scientific reports **3**: 1344.

Cardillo, A., M. Zanin, J. Gómez-Gardenes, M. Romance, A. J. G. del Amo and S. Boccaletti (2013). "Modeling the multi-layer nature of the European Air Transport Network: Resilience and passengers re-scheduling under random failures." The European Physical Journal Special Topics **215**(1): 23-33.

Carnap, R. (2014). Logical syntax of language, Routledge.

Carreras, B. A., V. E. Lynch, I. Dobson and D. E. Newman (2002). "Critical points and transitions in an electric power transmission model for cascading failure blackouts." Chaos An Interdisciplinary Journal of Nonlinear Science **12**(4): 985.

Cederman, L.-E. (2003). "Modeling the size of wars: from billiard balls to sandpiles." American Political Science Review **97**(1): 135-150.

Cellai, D., E. López, J. Zhou, J. P. Gleeson and G. Bianconi (2013). "Percolation in multiplex networks with overlap." Physical Review E **88**(5): 052811.

Chakrabarti, B. K., A. Chakraborti and A. Chatterjee (2007). Econophysics and sociophysics: trends and perspectives, John Wiley & Sons.

Chang, K.-C., K. Pearson and T. Zhang (2008). "Perron-Frobenius theorem for nonnegative tensors." Communications in Mathematical Sciences **6**(2): 507-520.

Checkland, P. (1983). "OR and the systems movement: mappings and conflicts." Journal of the Operational Research Society: 661-675.

Checkland, P. (1999). "Systems thinking, systems practice: includes a 30-year retrospective."

Cimenler, O., K. A. Reeves and J. Skvoretz (2014). "A regression analysis of researchers' social network metrics on their citation performance in a college of engineering." Journal of Informetrics **8**(3): 667-682.

Cobo, M. J., A. G. López-Herrera, E. Herrera-Viedma and F. Herrera (2011). "Science mapping software tools: Review, analysis, and cooperative study among tools." Journal of the Association for Information Science and Technology **62**(7): 1382-1402.

Cohen, E. B. and S. J. Lloyd (2014). "Disciplinary Evolution and the Rise of the Transdiscipline."

Collis, J. and R. Hussey (2013). Business research: A practical guide for undergraduate and postgraduate students, Palgrave macmillan.

Conway and Steward (2009). Managing and Shaping Innovation, Oxford University Press**:** 126-174.

Corominas-Murtra, B., B. Fuchs and S. Thurner (2014). "Detection of the elite structure in a virtual multiplex social system by means of a generalised K-core." PloS one **9**(12): e112606.

Corsi, M., C. D'Ippoliti and F. Lucidi (2010). "Pluralism at risk? Heterodox economic approaches and the evaluation of economic research in Italy." American Journal of Economics and Sociology **69**(5): 1495-1529.

Coscia, M., G. Rossetti, D. Pennacchioli, D. Ceccarelli and F. Giannotti (2013). "You know Because I Know": A multidimensional network approach to human resources problem. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, IEEE.

Cozzo, E., M. Kivelä, M. De Domenico, A. Solé, A. Arenas, S. Gómez, M. A. Porter and Y. Moreno (2013). "Clustering coefficients in multiplex networks." arXiv preprint arXiv:1307.6780.

Craven, P. and B. Wellman (1973). "The network city." Sociological inquiry **43**(3-4): 57-88.

Cummings, J. N. and S. Kiesler (2005). "Collaborative research across disciplinary and organizational boundaries." Social studies of science **35**(5): 703-722.

Daspit, J., T. C. Justice, N. G. Boyd and V. Mckee (2013). "Cross-functional team effectiveness: An examination of internal team environment, shared leadership, and cohesion influences." Team Performance Management: An International Journal **19**(1/2): 34-56.

David, N. (2013). Validating simulations. Simulating Social Complexity, Springer**:** 135-171.

Davidson, R. A. (2015). "Integrating disciplinary contributions to achieve community resilience to natural disasters." Civil Engineering and Environmental Systems **32**(1-2): 55-67.

Davis, J., A. MacDonald and L. White (2010). "Problem-structuring methods and project management: an example of stakeholder involvement using Hierarchical Process Modelling methodology." Journal of the Operational Research Society **61**(6): 893-904.

De Aguiar, M. and Y. Bar-Yam (2005). "Spectral analysis and the dynamic response of complex networks." Physical Review E **71**(1): 016106.

De Boer, Y., A. De Gier, M. Verschuur and B. De Wit (2006). "Building Bridges. Researchers on their experiences with interdisciplinary research in the Netherlands."

De Domenico, M. (2014). "MuxViz v0.2: visualization of multiplex networks

" Retrieved 17.06.2018, from http://muxviz.net, https://github.com/manlius/muxViz.

De Domenico, M., A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez and A. Arenas (2013). "Mathematical formulation of multilayer networks." Physical Review X **3**(4): 041022.

De Domenico, M., A. Solé-Ribalta, S. Gómez and A. Arenas (2014). "Navigability of interconnected networks under random failures." Proceedings of the National Academy of Sciences **111**(23): 8351-8356.

De Domenico, M., A. Solé-Ribalta, E. Omodei, S. Gómez and A. Arenas (2013). "Centrality in interconnected multilayer networks."

De Domenico, M., A. Solé-Ribalta, E. Omodei, S. Gómez and A. Arenas (2015). "Ranking in interconnected multilayer networks reveals versatile nodes." Nature communications **6**: 6868.

De Stefano, D., V. Fuccella, M. P. Vitale and S. Zaccarin (2013). "The use of different data sources in the analysis of co-authorship networks and scientific performance." Social Networks **35**(3): 370-381.

DEVELOPMENT, O. F. E. C.-O. A. (1996). THE KNOWLEDGE-BASED ECONOMY. Paris OCDE. **GD(96)102**.

Dictionary.com. (2018). "Methodology Definition." Retrieved 15.06.2018, from http://www.dictionary.com/browse/methodology?s=t.

Didelez, V. (2007). "Statistical causality." Consilience interdisplinary communications 2005 **2996**: 114-120.

Dixit, K., S. Kameshwaran, S. Mehta, V. Pandit and N. Viswanadham (2009). Towards simultaneously exploiting structure and outcomes in interaction networks for node ranking, IBM Research Report.

Dodds, P. S. and D. J. Watts (2005). "A generalized model of social and biological contagion." Journal of theoretical biology **232**(4): 587-604.

Dogan, M. and R. Pahre (1990). Creative marginality: Innovation at the intersections of social sciences, Westview Pr.

Dorogovtsev, S. N. and J. F. Mendes (2002). "Evolution of networks." Advances in physics **51**(4): 1079-1187.

Dunlavy, D. M., T. G. Kolda and W. P. Kegelmeyer (2011). Multilinear algebra for analyzing data with multiple linkages. Graph Algorithms in the Language of Linear Algebra, SIAM**:** 85-114.

Easterby-Smith, M., R. Thorpe, P. Jackson and A. Lowe (2008). "Management research (ed.)." London: SAGE. Ellison, N., Steinfeld, C., & Lampe, C.(2007). The benefits of Facebook" friends:" social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication **12**: 1143-1168.

Edquist, C. (1997). Systems of innovation: technologies, institutions, and organizations, Psychology Press.

Edquist, C. (2010). "Systems of innovation perspectives and challenges." African Journal of Science, Technology, Innovation and Development **2**(3): 14-45.

Egghe, L. (2006). "Theory and practise of the g-index." Scientometrics **69**(1): 131-152.

Einstein, A. (1918–1921). Induction and Deduction. Collected Papers of Albert Einstein, The Berlin Years: Writings. R. S. M. Janssen, et al. **Vol. 7**.

Eisenblätter, B., L. Santen, A. Schadschneider and M. Schreckenberg (1998). "Jamming transition in a cellular automaton model for traffic flow." Physical Review E **57**(2): 1309.

Eisenhardt, K. M. and B. N. Tabrizi (1995). "Accelerating adaptive processes: Product innovation in the global computer industry." Administrative science quarterly: 84-110.

Ellinas, C., M. Hall and A. Hultin (2014). DESIGN THROUGH FAILURE: A NETWORK PERSPECTIVE. DS 77: Proceedings of the DESIGN 2014 13th International Design Conference.

Ellison, N. B., J. Vitak, R. Gray and C. Lampe (2014). "Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes." Journal of Computer-Mediated Communication **19**(4): 855-870.

Emilio, F. and R. A. E. (2013). "Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index." Journal of the American Society for Information Science and Technology **64**(11): 2332-2339.

EPSRC (2014). Delivery Plan 2015/2016. Engineering and Physical Sciences Research Council, Polaris House, North Start Avenue, Swindon, SN2 1ET, EPSRC.

Epstein, J. M. (2008). "Why Model?" Journal of Artificial Societies and Social Simulation **11**(4): 12.

Erdős, P. and A. Rényi (1959). "On random graphs." Publicationes Mathematicae Debrecen **6**: 290-297.

Eskinasi, M., E. Rouwette and J. Vennix (2009). "Simulating urban transformation in Haaglanden, the Netherlands." System Dynamics Review **25**(3): 182-206.

Etzkowitz, H. (2008). The triple helix: university-industry-government innovation in action, Routledge.

Etzkowitz, H. and M. Ranga (2015). Triple Helix systems: an analytical framework for innovation policy and practice in the Knowledge Society. Entrepreneurship and Knowledge Exchange, Routledge**:** 117-158.

Euler, L. (1741). "Solutio problematis ad geometriam situs pertinentis." Commentarii academiae scientiarum Petropolitanae(8): 128-140.

Fabrikant, A., E. Koutsoupias and C. H. Papadimitriou (2002). Heuristically optimized trade-offs: A new paradigm for power laws in the Internet. International Colloquium on Automata, Languages, and Programming, Springer.

Farine, D. R. (2014). "Measuring phenotypic assortment in animal social networks: weighted associations are more robust than binary edges." Animal Behaviour **89**: 141-153.

Fleming, L., C. King and A. Juda (2007). "Small worlds and innovation." Organization Science **18**(6): 938-954.

Freeman, C. (2013). Economics of industrial innovation, Routledge.

Freeman, L., D. White and A. Romney (1992). Research methods in social network analysis, 1st, Transaction, New Brunswick.

Freeman, L. C. (1977). "A set of measures of centrality based on betweenness." Sociometry: 35-41.

Freeman, L. C. (1978). "Centrality in social networks conceptual clarification." Social networks **1**(3): 215-239.

Freeman, L. C. (1979). "Centrality in social networks: Conceptual clarification." Social Networks **1**(3): 215-239.

Freeman, L. C. (1980). "The gatekeeper, pair-dependency and structural centrality." Quality and Quantity **14**(4): 585-592.

Funk, S. and V. A. Jansen (2010). "Interacting epidemics on overlay networks." Physical Review E **81**(3): 036118.

Geurts, P., D. Ernst and L. Wehenkel (2006). "Extremely randomized trees." Machine learning **63**(1): 3-42.

Ghoshal, G., L. Chi and A.-L. Barabasi (2013). "Uncovering the role of elementary processes in network evolution." Sci. Rep. **3**.

Gomez, S., A. Diaz-Guilera, J. Gomez-Gardenes, C. J. Perez-Vicente, Y. Moreno and A. Arenas (2013). "Diffusion dynamics on multiplex networks." Physical review letters **110**(2): 028701.

Granell, C., S. Gómez and A. Arenas (2014). "Competing spreading processes on multiplex networks: awareness and epidemics." Physical review E **90**(1): 012808.

Granovetter, M. S. (1973). "The strength of weak ties." American journal of sociology: 1360-1380.

Greene, W. (2008). Econometric Analysis.

Greenland, S. (1998). "Induction versus Popper: substance versus semantics." International Journal of Epidemiology **27**(4): 543-548.

Guan, J. and N. Liu (2016). "Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy." Research policy **45**(1): 97-112.

Guan, J., K. Zuo, K. Chen and R. C. Yam (2016). "Does country-level R&D efficiency benefit from the collaboration network structure?" Research Policy **45**(4): 770-784.

Guimera, R., B. Uzzi, J. Spiro and L. A. N. Amaral (2005). "Team assembly mechanisms determine collaboration network structure and team performance." Science **308**(5722): 697-702.

Haight, F. A. (1967). Handbook of the Poisson distribution.

Hale, K. (2012). Collaboration in academic R&D: A decade of growth in pass-through funding. InfoBrief. NSF 12-325, National Science Foundation.

Hamill, J. T. (2006). Analysis of layered social networks, DTIC Document.

Hâncean, M.-G. and M. Perc (2016). "Homophily in coauthorship networks of East European sociologists." Scientific reports **6**: 36152.

Hargadon, A. and R. I. Sutton (1997). "Technology brokering and innovation in a product development firm." Administrative science quarterly: 716-749.

Harzing, A.-W. (2008). "Reflections on the h-index."

Haskel, J. and G. Wallis (2013). "Public support for innovation, intangible investment and productivity growth in the UK market sector." Economics Letters **119**(2): 195-198.

Hayes, J. F. and T. V. G. Babu (2004). Modeling and analysis of telecommunications networks, John Wiley & Sons.

Heaney, M. T. (2014). "Multiplex networks and interest group influence reputation: An exponential random graph model." Social Networks **36**: 66-81.

Henderson, L. (2018). The Problem of Induction. The Stanford Encyclopedia of Philosophy. E. N. Zalta. **Summer 2018 Edition**.

Hirsch, J. E. (2007). "Does the h index have predictive power?" Proceedings of the National Academy of Sciences **104**(49): 19193-19198.

Ho, T. M., H. V. Nguyen, T.-T. Vuong, Q.-M. Dam, H.-H. Pham and Q.-H. Vuong (2017). "Exploring Vietnamese co-authorship patterns in social sciences with basic network measures of 2008-2017 Scopus data." F1000Research **6**.

Hollaender, K., M. C. Loibl and A. Wilts (2008). Management. Handbook of Transdisciplinary Research**:** 385-397.

Holme, P. and J. Saramäki (2013). Temporal networks, Springer.

Horwitz, S. K. (2005). "The compositional impact of team diversity on performance: Theoretical considerations." Human resource development review **4**(2): 219-245.

Hufnagel, L., D. Brockmann and T. Geisel (2004). "Forecast and control of epidemics in a globalized world." Proceedings of the National Academy of Sciences of the United States of America **101**(42): 15124-15129.

Hume, D. (2003). A treatise of human nature, Courier Corporation.

Huutoniemi, K. and I. Rafols (2016). "Interdisciplinarity in research evaluation."

Innovation, U. R. a. (2018). "Global Challenges Research Fund (GCRF)." Retrieved 18.06.2018, from https://www.ukri.org/research/global-challenges-research-fund/.

Ito, T., T. Kaneta and S. Sundstrom (2015). "Does university entrepreneurship work in Japan?: a comparison of industry-university research funding and technology transfer activities between the UK and Japan." Journal of Innovation and Entrepreneurship **5**(1): 8.

Jackson, M. O. (2010). Social and economic networks, Princeton university press.

Jackson, S. L. (2014). Research methods: A modular approach, Cengage Learning.

Jacobs, J. A. and S. Frickel (2009). "Interdisciplinarity: A critical assessment." Annual review of Sociology **35**: 43-65.

Josheski, D. and C. Koteski (2011). "The causal relationship between patent growth and growth of GDP with quarterly data in the G7 countries: cointegration, ARDL and error correction models."

Junjie, P. and W. Dingwei (2006). An ant colony optimization algorithm for multiple travelling salesman problem. First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06), IEEE.

Kachra, A. and R. E. White (2008). "Know-how transfer: the role of social, economic/competitive, and firm boundary factors." Strategic Management Journal **29**(4): 425-445.

Kang, Y.-B., Y.-F. Li and R. L. Coppel (2015). Capturing researcher expertise through mesh classification. Proceedings of the 8th International Conference on Knowledge Capture, ACM.

Karlebach, G. and R. Shamir (2008). "Modelling and analysis of gene regulatory networks." Nature Reviews Molecular Cell Biology **9**(10): 770.

Karsai, M., K. Kaski, A.-L. Barabási and J. Kertész (2012). "Universal features of correlated bursty behaviour." Scientific reports **2**.

Katz, J. S. and B. R. Martin (1997). "What is research collaboration?" Research Policy **26**(1): 1-18.

Kaur, J. and V. Gupta (2010). "Effective approaches for extraction of keywords." International Journal of Computer Science Issues **7**(6): 144-148.

Keener, J. P. (1993). "The Perron–Frobenius theorem and the ranking of football teams." SIAM review **35**(1): 80-93.

Keller, R. T. (2001). "Cross-functional project groups in research and new product development: Diversity, communications, job stress, and outcomes." Academy of management journal **44**(3): 547-555.

Kephart, J. O. (1994). How topology affects population dynamics. SANTA FE INSTITUTE STUDIES IN THE SCIENCES OF COMPLEXITY-PROCEEDINGS VOLUME-, ADDISON-WESLEY PUBLISHING CO.

Khan, G. F. and H. W. Park (2013). "The e-government research domain: A triple helix network analysis of collaboration at the regional, country, and institutional levels." Government Information Quarterly **30**(2): 182-193.

Kim, J. Y. and K.-I. Goh (2013). "Coevolution and correlated multiplexity in multiplex networks." Physical review letters **111**(5): 058702.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno and M. A. Porter (2014). "Multilayer networks." Journal of complex networks **2**(3): 203-271.

Klein, J. T. (2008). "Evaluation of interdisciplinary and transdisciplinary research." American journal of preventive medicine **35**(2): S116-S123.

Klein, O., Y. Grignon, C. Pinelli, T. Civit, J. Auque and J. C. Marchal (2004). "Pleomorphic xanthoastrocytoma. A review of five observations." Neurochirurgie **50**(5): 515-520.

Kleinberg, J. M. (1999). "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) **46**(5): 604-632.

Kleinberg, J. M., R. Kumar, P. Raghavan, S. Rajagopalan and A. S. Tomkins (1999). The web as a graph: Measurements, models, and methods. International Computing and Combinatorics Conference, Springer.

Kline, S. J. and N. Rosenberg (2010). An overview of innovation. Studies On Science And The Innovation Process: Selected Works of Nathan Rosenberg, World Scientific**:** 173-203.

Ko, Y. (2012). A study of term weighting schemes using class information for text classification. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Portland, Oregon, USA, ACM**:** 1029-1030.

Kohler, U. and F. Kreuter (2012). Data Analysis Using Stata, Stata Press.

Kolda, T. and B. Bader (2006). The TOPHITS model for higher-order web link analysis. Workshop on link analysis, counterterrorism and security.

Kolda, T. G., B. W. Bader and J. P. Kenny (2005). Higher-order web link analysis using multilinear algebra. Data Mining, Fifth IEEE International Conference on, IEEE.

König, B., K. Diehl, K. Tscherning and K. Helming (2013). "A framework for structuring interdisciplinary research management." Research Policy **42**(1): 261-272.

Kossiakoff, A., W. N. Sweet, S. Seymour and S. M. Biemer (2011). Systems engineering principles and practice, John Wiley & Sons.

Krackhardt, D. (1987). "Cognitive social structures." Social networks **9**(2): 109-134.

Krapivsky, P. L., S. Redner and F. Leyvraz (2000). "Connectivity of growing random networks." Physical review letters **85**(21): 4629.

Kreye, M. E. (2011). Uncertainty analysis in competitive bidding for service contracts, University of Bath.

Kronegger, L., A. Ferligoj and P. Doreian (2011). "On the dynamics of national scientific systems." Quality and Quantity **45**(5): 989-1015.

Kueffer, C. and G. H. Hadorn (2008). "How to achieve effectiveness in problem-oriented landscape research: the example of research on biotic invasions." Living reviews in landscape research **2**.

Kuhn, T. S. (2012). The structure of scientific revolutions, University of Chicago press.

Kumar, S. and P. Phrommathed (2005). Research methodology, Springer.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes: Ed by John Worrall and Gregory Currie, Cambridge University Press.

Lambiotte, R. and M. Rosvall (2012). "Ranking and clustering of nodes in networks with smart teleportation." Physical Review E **85**(5): 056107.

Lane, P. J. and M. Lubatkin (1998). "Relative absorptive capacity and interorganizational learning." Strategic management journal: 461-477.

Larsen, L., C. Thomas, M. Eppinga and T. Coulthard (2014). "Exploratory modeling: Extracting causality from complexity." Eos, Transactions American Geophysical Union **95**(32): 285-286.

Lawrence, P. R. and J. W. Lorsch (1986). "Organization and environment: managing differentiation and integration (Harvard Business School Classics)."

Leavy, P. (2011). Oral history: Understanding qualitative research, Oxford University Press.

Lee, K.-M., J. Y. Kim, W.-k. Cho, K.-I. Goh and I. Kim (2012). "Correlated multiplexity and connectivity of multiplex random networks." New Journal of Physics **14**(3): 033027.

Lee, S. and B. Bozeman (2005). "The impact of research collaboration on scientific productivity." Social Studies of Science **35**(5): 673-702.

Lehmann, S., A. D. Jackson and B. E. Lautrup (2006). "Measures for measures." Nature **444**(7122): 1003.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance and M. Hurst (2007). "Cascading behavior in large blog graphs." arXiv preprint arXiv:0704.2803.

Leskovec, J., A. Singh and J. Kleinberg (2006). Patterns of influence in a recommendation network. Advances in Knowledge Discovery and Data Mining, Springer**:** 380-389.

Levitt, J. M. and M. Thelwall (2008). "Is multidisciplinary research more highly cited? A macrolevel study." Journal of the American Society for Information Science and Technology **59**(12): 1973-1984.

Leydesdorff, L., P. Nightingale, A. O'Hare, I. Rafols and A. Stirling (2011). How rankings can suppress interdisciplinarity. The case of innovation studies and business and management, Georgia Institute of Technology.

Li, E. Y., C. H. Liao and H. R. Yen (2013). "Co-authorship networks and research impact: A social capital perspective." Research Policy **42**(9): 1515-1530.

Li, H. and Y. Zhang (2007). "The role of managers' political networking and functional experience in new venture performance: Evidence from China's transition economy." Strategic management journal **28**(8): 791-804.

Linden, R., L. F. Barbosa and L. A. Digiampietri (2017). "“Brazilian style science”—an analysis of the difference between Brazilian and international Computer Science departments and graduate programs using social networks analysis and bibliometrics." Social Network Analysis and Mining **7**(1): 44.

Lundvall, B.-ä. and B. Johnson (1994). "The Learning Economy." Journal of Industry Studies **1**(2): 23-42.

Lundvall, B. Å. (1998). "Why study national systems and national styles of innovation?" Technology Analysis and Strategic Management **10**(4): 407-421.

Lundvall, B. Å. (2007). "National innovation systems - Analytical concept and development tool." Industry and Innovation **14**(1): 95-119.

Lundvall, B. Å., B. Johnson, E. S. Andersen and B. Dalum (2002). "National systems of production, innovation and competence building." Research Policy **31**(2): 213-231.

Lusher, D., J. Koskinen and G. Robins (2013). Exponential random graph models for social networks: Theory, methods, and applications, Cambridge University Press.

Luukkonen, T., O. Persson and G. Sivertsen (1992). "Understanding Patterns of International Scientific Collaboration." Science, Technology & Human Values **17**(1): 101-126.

Luukkonen, T., R. J. W. Tijssen, O. Persson and G. Sivertsen (1993). "The measurement of international scientific collaboration." Scientometrics **28**(1): 15-36.

Lytras, M. D., P. O. De Pablos, A. Ziderman, A. Roulstone, H. Maurer and J. B. Imber (2010). Knowledge Management, Information Systems, E-Learning, and Sustainability Research: Third World Summit on the Knowledge Society, WSKS 2010, Corfu, Greece, September 22-24, 2010, Proceedings, Springer.

Magnani, M. and L. Rossi (2013). Pareto distance for multi-layer network analysis. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer.

Maher, P. (2010). Lecture 21

Popper on Auxiliary Hypotheses.

Manring, S. L. (2014). "The role of universities in developing interdisciplinary action research collaborations to understand and manage resilient social-ecological systems." Journal of Cleaner Production **64**: 125-135.

Mansilla, V. B., M. Lamont and K. Sato (2013). "Successful interdisciplinary collaborations: The contributions of shared socio-emotional-cognitive platforms to interdisciplinary synthesis."

Marcella, C., D. I. Carlo and L. Federico (2010). "Pluralism at Risk? Heterodox Economic Approaches and the Evaluation of Economic Research in Italy." American Journal of Economics and Sociology **69**(5): 1495-1529.

Marro, J. and R. Dickman (2005). Nonequilibrium phase transitions in lattice models, Cambridge University Press.

Martin, B. R. and R. Whitley (2010). "The UK Research Assessment Exercise: A case of regulatory capture?" Reconfiguring Knowledge Production: Changing Authority Relationships in the Sciences and Their Consequences for Intellectual Innovation: 51-80.

Maurer, M. S. (2007). Structural Awareness in Complex Product Design. Doktor-Ingenieurs, Technischen Universität München.

May, R. M. and A. L. Lloyd (2001). "Infection dynamics on scale-free networks." Physical Review E **64**(6): 066112.

McCain, K. W. (1998). "Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field." Scientometrics **41**(3): 389-410.

McCarty, C., J. W. Jawitz, A. Hopkins and A. Goldman (2013). "Predicting author h-index using characteristics of the co-author network." Scientometrics **96**(2): 467-483.

McFadyen, M. A. and A. A. Cannella (2004). "Social capital and knowledge creation: Diminishing returns of the number and strength of exchange relationships." Academy of Management Journal **47**(5): 735-746.

McFadyen, M. A. and A. A. Cannella (2005). "Knowledge creation and the location of university research scientists' interpersonal exchange relations: Within and beyond the university." Strategic Organization **3**(2): 131-155.

McFadyen, M. A., M. Semadeni and A. A. Cannella Jr (2009). "Value of strong ties to disconnected others: Examining knowledge creation in biomedicine." Organization science **20**(3): 552-564.

McTear, M., Z. Callejas and D. Griol (2016). "The conversational interface." Springer **6**(94): 102.

Mehmood, A., G. S. Choi, O. F. von Feigenblatt and H. W. Park (2016). "Proving ground for social network analysis in the emerging research area "Internet of Things"(IoT)." Scientometrics **109**(1): 185-201.

Melin, G. (2000). "Pragmatism and self-organization: Research collaboration on the individual level." Research Policy **29**(1): 31-40.

Melnik, S., M. A. Porter, P. J. Mucha and J. P. Gleeson (2014). "Dynamics on modular networks with heterogeneous correlations." Chaos: An Interdisciplinary Journal of Nonlinear Science **24**(2): 023106.

Mena-Chalco, J. P., L. A. Digiampietri, F. M. Lopes and R. M. Cesar (2014). "Brazilian bibliometric coauthorship networks." Journal of the Association for Information Science and Technology **65**(7): 1424-1445.

Mendenhall, W. M. and T. L. Sincich (2016). Statistics for Engineering and the Sciences, Chapman and Hall/CRC.

Menichetti, G., D. Remondini, P. Panzarasa, R. J. Mondragón and G. Bianconi (2014). "Weighted multiplex networks." PloS one **9**(6): e97857.

Merz, B., J. Friedrich, M. Disse, J. Schwarz, J. G. Goldammer and J. Wächter (2006). "Possibilities and limitations of interdisciplinary, user-oriented research: experiences from the German Research Network Natural Disasters." Natural hazards **38**(1-2): 3-20.

Metke-Jimenez, A. and S. Karimi (2015). "Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms." arXiv preprint arXiv:1504.06936.

Michalisin, M. D., S. J. Karau and C. Tangpong (2007). "Leadership's activation of team cohesion as a strategic asset: An empirical simulation." Journal of Business Strategies **24**(1): 1.

Mihalcea, R. and P. Tarau (2004). Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing.

Milgram, S. (1967). "The small world problem." Psychology today **2**(1): 60-67.

Min, B., S. Do Yi, K.-M. Lee and K.-I. Goh (2014). "Network robustness of multiplex networks with interlayer degree correlations." Physical Review E **89**(4): 042811.

Mingers, J. (2011). "Soft OR comes of age—but not everywhere!" Omega **39**(6): 729-741.

Molas-Gallart, J. and A. Salter (2002). "Diversity and excellence: considerations on research policy." IPTS report **66**(5): 5-13.

Moreno, J. L. and H. H. Jennings (1934). "Who shall survive?".

Morris, R. G. and M. Barthelemy (2012). "Transport on coupled spatial networks." Physical review letters **109**(12): 128703.

Morris, R. G. and M. Barthelemy (2014). Spatial effects: Transport on interdependent networks. Networks of networks: the last frontier of complexity, Springer**:** 145-161.

Motter, A. E. and Y.-C. Lai (2002). "Cascade-based attacks on complex networks." Physical Review E **66**(6): 065102.

Mucha, P. J. and M. A. Porter (2010). "Communities in multislice voting networks." Chaos: An Interdisciplinary Journal of Nonlinear Science **20**(4): 041108.

Mujtaba, M. (1994). "Simulation modelling of a manufacturing enterprise with complex material, information and control flows." International Journal of Computer Integrated Manufacturing **7**(1): 29-46.

Munoz, D. A., J. P. Queupil and P. Fraser (2016). "Assessing collaboration networks in educational research: A co-authorship-based social network analysis approach." International Journal of Educational Management **30**(3): 416-436.

Nahapiet, J. and S. Ghoshal (1998). "Social capital, intellectual capital and the organizational advantage Academy of Management Review 23 (2): 242–266." CrossRef Google Scholar.

National Academies of Sciences, E. and Medicine (2005). Facilitating interdisciplinary research, Washington, DC: National Academies Press.

Nelson, R. R. (2009). An evolutionary theory of economic change, harvard university press.

Nerkar, A. and S. Paruchuri (2005). "Evolution of R&D capabilities: The role of knowledge networks within a firm." Management science **51**(5): 771-785.

Newman, M. (2010). Networks: an introduction, Oxford university press**:** 1-12.

Newman, M. E. (2001). "The structure of scientific collaboration networks." Proceedings of the National Academy of Sciences **98**(2): 404-409.

Newman, M. E. (2003). "The structure and function of complex networks." SIAM review **45**(2): 167-256.

Newman, M. E. (2006). "Modularity and community structure in networks." Proceedings of the National Academy of Sciences **103**(23): 8577-8582.

Newman, M. E. J. and M. Girvan (2004). "Finding and evaluating community structure in networks." Physical Review E **69**(2): 026113.

Nicosia, V., G. Bianconi, V. Latora and M. Barthelemy (2013). "Growing multiplex networks." Physical review letters **111**(5): 058701.

Nicosia, V., G. Bianconi, V. Latora and M. Barthelemy (2014). "Nonlinear growth and condensation in multiplex networks." Physical Review E **90**(4): 042807.

Nicosia, V. and V. Latora (2015). "Measuring and modeling correlations in multiplex networks." Physical Review E **92**(3): 032805.

Nonaka, I., P. Byosiere, C. C. Borucki and N. Konno (1994). "Organizational knowledge creation theory: a first comprehensive test." International Business Review **3**(4): 337-351.

Nooteboom, B. (2000). "Learning by interaction: absorptive capacity, cognitive distance and governance." Journal of management and governance **4**(1-2): 69-92.

Nunkoo, R., D. Gursoy and H. Ramkissoon (2013). "Developments in hospitality marketing and management: Social network analysis and research themes." Journal of Hospitality Marketing & Management **22**(3): 269-288.

Obstfeld, M. (2015). "Trilemmas and trade-offs: living with financial globalisation."

Ogata, K. (2002). Modern control engineering, Prentice hall India.

Opsahl, T. (2009). Structure and evolution of weighted networks, Queen Mary, University of London.

Opsahl, T., F. Agneessens and J. Skvoretz (2010). "Node centrality in weighted networks: Generalizing degree and shortest paths." Social networks **32**(3): 245-251.

Opsahl, T. and P. Panzarasa (2009). "Clustering in weighted networks." Social networks **31**(2): 155-163.

Ormerod, P. and B. Rosewell (2009). Validation and verification of agent-based models in the social sciences. Epistemological aspects of computer simulation in the social sciences, Springer**:** 130-140.

Ortega, J. L. (2014). "Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search." Journal of Informetrics **8**(3): 728-737.

Oxford-Dictionaries (2017). Discipline. English Oxford Dictionaries, Oxford Dictionaries.

Padgett, J. F. and C. K. Ansell (1993). "Robust Action and the Rise of the Medici, 1400-1434." American journal of sociology **98**(6): 1259-1319.

Page, L., S. Brin, R. Motwani and T. Winograd (1999). "The PageRank citation ranking: bringing order to the web."

Page, S. E. (2007). The difference. How the power of diversity creates better groups, firms, schools, and societies. Princenton, NJ, Princeton University Press.

Pahwa, S., M. Youssef and C. Scoglio (2014). Electrical networks: an introduction. Networks of networks: the last frontier of complexity, Springer**:** 163-186.

Pan, R. K. and J. Saramäki (2012). "The strength of strong ties in scientific collaboration networks." EPL (Europhysics Letters) **97**(1): 18007.

Parameswaran, A., H. Garcia-Molina and A. Rajaraman (2010). "Towards the web of concepts: Extracting concepts from large datasets." Proceedings of the VLDB Endowment **3**(1-2): 566-577.

Parandehgheibi, M. and E. Modiano (2013). Robustness of interdependent networks: The case of communication networks and the power grid. Global Communications Conference (GLOBECOM), 2013 IEEE, IEEE.

Parker, G. M. (2003). Cross-functional teams: Working with allies, enemies, and other strangers, John Wiley & Sons.

Pastor-Satorras, R. and A. Vespignani (2001). "Epidemic spreading in scale-free networks." Physical review letters **86**(14): 3200.

Perry-Smith, J. E. (2006). "Social yet creative: The role of social relationships in facilitating individual creativity." Academy of Management journal **49**(1): 85-101.

Peters, T. W. and R. Waterman Jr "R.(1982)." Search of Excellence.

Phelps, C., R. Heidl and A. Wadhwa (2012). "Knowledge, networks, and knowledge networks: A review and research agenda." Journal of management **38**(4): 1115-1166.

Phillips, N. (2010). The slow death of pluralism. International Political Economy, Routledge**:** 84-92.

Porter, M. (accessed 2018). "Snowball." Retrieved 17.06.2018, from http://snowballstem.org/.

Qi, L. (2005). "Eigenvalues of a real supersymmetric tensor." Journal of Symbolic Computation **40**(6): 1302-1324.

Quinn, R. E. (1988). Beyond rational management: Mastering the paradoxes and competing demands of high performance, Jossey-Bass.

Raasch, C., V. Lee, S. Spaeth and C. Herstatt (2013). "The rise and fall of interdisciplinary research: The case of open source innovation." Research policy **42**(5): 1138-1151.

Radicchi, F. (2014). "Driving interconnected networks to supercriticality." Physical Review X **4**(2): 021014.

Rafols, I. (2007). "Strategies for knowledge acquisition in bionanotechnology: Why are interdisciplinary practices less widespread than expected?" Innovation **20**(4): 395-412.

Rafols, I., L. Leydesdorff, A. O'Hare, P. Nightingale and A. Stirling (2012). "How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management." Research Policy **41**(7): 1262-1282.

Rafols, I. and M. Meyer (2007). "How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors." Scientometrics **70**(3): 633-650.

RCUK. (2017). Retrieved 05.09.2017, 2017, from http://www.rcuk.ac.uk/about/aims-and-organisation/.

Reagans, R. and B. McEvily (2003). "Network structure and knowledge transfer: The effects of cohesion and range." Administrative Science Quarterly **48**(2): 240-267.

Repko, A. F. (2008). Interdisciplinary research: Process and theory, Sage.

Rigby, D. K., J. Sutherland and H. Takeuchi (2016). "Embracing agile." Harvard Business Review **94**(5): 40-50.

Rittel, H. W. and M. M. Webber (1973). "Dilemmas in a general theory of planning." Policy sciences **4**(2): 155-169.

Robson, C. (2002). "Real world research. 2nd." Edition. Blackwell Publishing. Malden.

Roche, A. J. and L. N. Rickard (2017). "Cocitation or Capacity-Building? Defining Success within an Interdisciplinary, Sustainability Science Team." Frontiers in Communication **2**: 13.

Rocklin, M. and A. Pinar (2013). "On clustering on graphs with multiple edge types." Internet Mathematics **9**(1): 82-112.

Rodrigues, A. and J. Bowers (1996). "System dynamics in project management: a comparative analysis with traditional methods." System Dynamics Review **12**(2): 121-139.

Rogers, E. M. and D. K. Bhowmik (1970). "Homophily-heterophily: Relational concepts for communication research." Public opinion quarterly **34**(4): 523-538.

Rombach, P., M. A. Porter, J. H. Fowler and P. J. Mucha (2017). "Core-periphery structure in networks (revisited)." SIAM Review **59**(3): 619-646.

Rose, S., D. Engel, N. Cramer and W. Cowley (2010). "Automatic keyword extraction from individual documents." Text Mining: Applications and Theory: 1-20.

Rouwette, E. A., J. A. Vennix and T. v. Mullekom (2002). "Group model building effectiveness: a review of assessment studies." System Dynamics Review **18**(1): 5-45.

Ruane, F. and R. S. J. Tol (2008). "Rational (successive) h-indices: An application to economics in the Republic of Ireland." Scientometrics **75**(2): 395-405.

Saavedra, S., F. Reed-Tsochas and B. Uzzi (2008). "Asymmetric disassembly and robustness in declining networks." Proceedings of the National Academy of Sciences **105**(43): 16466-16471.

Sahneh, F. D. and C. Scoglio (2014). "Competitive epidemic spreading over arbitrary multilayer networks." Physical Review E **89**(6): 062817.

Sahneh, F. D., C. Scoglio and F. N. Chowdhury (2013). Effect of coupling on the epidemic threshold in interconnected complex networks: A spectral analysis. American Control Conference (ACC), 2013, IEEE.

Sánchez-García, R. J., E. Cozzo and Y. Moreno (2014). "Dimensionality reduction and spectral properties of multilayer networks." Physical Review E **89**(5): 052815.

Sargent, R. G. (2001). Verification and validation: some approaches and paradigms for verifying and validating simulation models. Proceedings of the 33nd conference on Winter simulation, IEEE Computer Society.

Sargent, R. G. (2013). "Verification and validation of simulation models." Journal of Simulation **7**(1): 12-24.

Sarigöl, E., R. Pfitzner, I. Scholtes, A. Garas and F. Schweitzer (2014). "Predicting scientific success based on coauthorship networks." EPJ Data Science **3**(1): 9.

Saunders, M., P. Lewis and A. Thornhill (2011). Research methods for business students, 5/e, Pearson Education India.

Schelling, T. C. (1969). "Models of segregation." The American Economic Review **59**(2): 488-493.

Schilling, M. A. and C. C. Phelps (2007). "Interfirm collaboration networks: The impact of large-scale network structure on firm innovation." Management science **53**(7): 1113-1126.

Schwartz, K. and J.-T. Vilquin (2003). "Building the translational highway: toward new partnerships between academia and the private sector." Nature medicine **9**(5): 493.

Senge, P. M. (1991). "The fifth discipline, the art and practice of the learning organization." Performance+ Instruction **30**(5): 37-37.

Siedlok, F. and P. Hibbert (2014). "The Organization of Interdisciplinary Research: Modes, Drivers and Barriers." International Journal of Management Reviews **16**(2): 194-210.

Simons, T., L. H. Pelled and K. A. Smith (1999). "Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams." Academy of management journal **42**(6): 662-673.

Singh, J. (2005). "Collaborative networks as determinants of knowledge diffusion patterns." Management science **51**(5): 756-770.

Singh, J. and L. Fleming (2010). "Lone inventors as sources of breakthroughs: Myth or reality?" Management science **56**(1): 41-56.

Smola, A. J. and B. Schölkopf (2004). "A tutorial on support vector regression." Statistics and computing **14**(3): 199-222.

Solá, L., M. Romance, R. Criado, J. Flores, A. García del Amo and S. Boccaletti (2013). "Eigenvector centrality of nodes in multiplex networks." Chaos: An Interdisciplinary Journal of Nonlinear Science **23**(3): 033131.

Solé-Ribalta, A., M. De Domenico, S. Gómez and A. Arenas (2014). Centrality rankings in multiplex networks. Proceedings of the 2014 ACM conference on Web science, ACM.

Sole-Ribalta, A., M. De Domenico, N. E. Kouvaris, A. Diaz-Guilera, S. Gomez and A. Arenas (2013). "Spectral properties of the Laplacian of multiplex networks." Physical Review E **88**(3): 032807.

Solé-Ribalta, A., S. Gómez and A. Arenas (2016). "Congestion induced by the structure of multiplex networks." Physical review letters **116**(10): 108701.

Sonnenwald, D. H. (2007). "Scientific collaboration." Annual review of information science and technology **41**(1): 643-681.

Sparrowe, R. T., R. C. Liden, S. J. Wayne and M. L. Kraimer (2001). "Social networks and the performance of individuals and groups." Academy of management journal **44**(2): 316-325.

Sterman, J. (2000). Business dynamics, Irwin-McGraw-Hill.

Steve Conway and F. Steward (2009). Managing and Shaping Innovation, Oxford University Press**:** 126-174.

Stirling, A. (1998). "On the economics and analysis of diversity." Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper **28**: 1-156.

Stirling, A. (2007). "A general framework for analysing diversity in science, technology and society." Journal of the Royal Society Interface **4**(15): 707-719.

Stock, J. H. and M. W. Watson (2015). "Introduction to Econometrics (3rd Updated Edition)." Age (X3) **3**: 0.22.

Sun, Y., J. Han, X. Yan, P. S. Yu and T. Wu (2011). "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks." Proceedings of the VLDB Endowment **4**(11): 992-1003.

Szell, M. and S. Thurner (2010). "Measuring social dynamics in a massive multiplayer online game." Social networks **32**(4): 313-329.

Taebi, B., A. Correlje, E. Cuppen, M. Dignum and U. Pesch (2014). "Responsible innovation as an endorsement of public values: The need for interdisciplinary research." Journal of Responsible Innovation **1**(1): 118-124.

Taylor, J. (2011). "The Assessment of Research Quality in UK Universities: Peer Review or Metrics?" British Journal of Management **22**(2): 202-217.

Thornton, S. (2017). Karl Popper. The Stanford Encyclopedia of Philosophy. E. N. Zalta.

Tichy, N. M., M. L. Tushman and C. Fombrun (1979). "Social network analysis for organizations." Academy of management review **4**(4): 507-519.

Tipping, M. E. and C. M. Bishop (1999). "Probabilistic principal component analysis." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3): 611-622.

Tsai, W. and C. H. Wu (2010). "From the editors knowledge combination: A cocitation analysis." Academy of Management Journal **53**(3): 441-450.

Uddin, S., L. Hossain and K. Rasmussen (2013). "Network effects on scientific collaborations." PloS one **8**(2): e57546.

Utterback, J. M. and W. J. Abernathy (1975). "A dynamic model of process and product innovation." Omega **3**(6): 639-656.

Van Noorden, R. (2015). "Interdisciplinary research by the numbers." Nature News **525**(7569): 306.

Van Rijnsoever, F. J. and L. K. Hessels (2011). "Factors associated with disciplinary and interdisciplinary research collaboration." Research policy **40**(3): 463-472.

Vennix, J. A. (1999). "Group model-building: tackling messy problems." System Dynamics Review **15**(4): 379-401.

Voit, E. O. and T. Radivoyevitch (2000). "Biochemical systems analysis of genome-wide expression data." Bioinformatics **16**(11): 1023-1037.

Volz, E. (2008). "SIR dynamics in random networks with heterogeneous connectivity." Journal of mathematical biology **56**(3): 293-310.

von Bohlen und Halbach, O. (2011). "How to judge a book by its cover? How useful are bibliometric indices for the evaluation of “scientific quality” or “scientific productivity”?" Annals of Anatomy - Anatomischer Anzeiger **193**(3): 191-196.

Von Luxburg, U. (2007). "A tutorial on spectral clustering." Statistics and computing **17**(4): 395-416.

Wagner, C. S., J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols and K. Börner (2011). "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature." Journal of informetrics **5**(1): 14-26.

Wang, H., Q. Li, G. D'Agostino, S. Havlin, H. E. Stanley and P. Van Mieghem (2013). "Effect of the interconnected network structure on the epidemic threshold." Physical Review E **88**(2): 022801.

Wang, P., G. Robins, P. Pattison and E. Lazega (2013). "Exponential random graph models for multilevel networks." Social Networks **35**(1): 96-115.

Wasserman, S. and K. Faust (1994). Social network analysis: Methods and applications, Cambridge university press.

Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of 'small-world'networks." nature **393**(6684): 440-442.

White, H. D., B. Wellman and N. Nazer (2004). "Does citation reflect social structure?: Longitudinal evidence from the "Globenet" interdisciplinary research group." Journal of the American Society for information Science and Technology **55**(2): 111-126.

Whitley, R. (2000). The intellectual and social organization of the sciences, Oxford University Press on Demand.

Wikipedia. "Outline of academic disciplines." Retrieved 21.04.2017 at 02:54, from https://en.wikipedia.org/wiki/Outline_of_academic_disciplines.

Wilensky, U. and W. Rand (2015). An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo, MIT Press.

Willmott, H. (2011). "Listing perilously." Organization-Interdisc Journ of Organiz Theory and Society **18**(4): 447.

Wu, Y. and Z. Duan (2015). "Social network analysis of international scientific collaboration on psychiatry research." International journal of mental health systems **9**(1): 2.

Xiang, X., R. Kennedy, G. Madey and S. Cabaniss (2005). Verification and validation of agent-based scientific simulation models. Agent-directed simulation conference.

Yan, X., L. Zhai and W. Fan (2013). "C-index: A weighted network node centrality measure for collaboration competence." Journal of Informetrics **7**(1): 223-239.

Ye, Q., T. Li and R. Law (2013). "A coauthorship network analysis of tourism and hospitality research collaboration." Journal of Hospitality & Tourism Research **37**(1): 51-76.

Yegros-Yegros, A., I. Rafols and P. D'Este (2015). "Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity." PloS one **10**(8): e0135095.

Yong, A. (2014). "Critique of Hirsch's citation index: A combinatorial Fermi problem." Notices of the AMS **61**(9): 1040-1050.

Zachary, W. W. (1977). "An information flow model for conflict and fission in small groups." Journal of anthropological research **33**(4): 452-473.

Zaltman, G. and R. Duncan (1977). Strategies for planned change, Wiley.

Zaremba, A. and T. Aste (2014). "Measures of causality in complex datasets with application to financial data." Entropy **16**(4): 2309-2349.

Zeigler, E. F. (1990). "Don't forget the profession when choosing a name." The Evolving Undergraduate Major. Champaign: Human Kinetics.

Zhao, S. X., R. Rousseau and F. Y. Ye (2011). "h-Degree as a basic measure in weighted networks." Journal of Informetrics **5**(4): 668-677.

Zhou, J., S. J. Shin, D. J. Brass, J. Choi and Z.-X. Zhang (2009). "Social networks, personal values, and creativity: evidence for curvilinear and interaction effects." Journal of applied psychology **94**(6): 1544.

Zurutuza, A. and C. Marinelli (2014). "Challenges and opportunities in graphene commercialization." Nature nanotechnology **9**(10): 730.

Appendix A: Content-based approach Scopus search-terms

Arts	european art music satellit televis signal moscow film school salari fashion design popular film industri interior design servic digit televis broadcast instrument digit interfac music instrument digit american cultur danc digit imag sensor independ industri design physic public domain end fashion design color televis set	Biology	differ integr membran data analyt method normal circadian rhythm perpetu immun system landscap limnolog framework malthusian natur select frozen section procedur basic marin scienc essenti fatti acid central nervous system structur align program genet code expans infer evolutionari relationship undifferenti biolog cell human bodi function
Chemical Engineering	organ small molecul tradit inorgan electron modern petroleum refineri simpl molecular machin mani unit oper process engin focus contract packag engin propos devic abl complex molecular machin function specif consensus modern materi scienc event molecular nanotechnology chemic reaction engin nation nanotechnolog initi	Chemistry	combinatori chemistri librari initio quantum chemistri constant chemic composit specif heat capac quantum mani bodi modern organ synthesi individu electrod potenti appar stimul emiss complex chemic reaction natur product compound radioact decay process measur electrod potenti term exotherm process analyt chemistri studi solid state inorgan
Civil Engineering	advanc structur analysi critic state soil intellig transport societi land survey system feder mine safeti structur system transfer moh scale hard former american congress steel industri wastewat wind wave system earthquak engin research geotechn engin project finit element concept rectangular survey system critic state concept	Computer Science	artifici intellig comput theori fail oper system memori manag techniqu analog electron circuit cyber secur linear tempor logic electr comprehens oper dynam neural network new comput architectur dynam data structur binari merg algorithm intellig agent paradigm ture machin model communic protocol standard

Economics	human poverti index statist mechan descript theoret market structur ration util maxim feminist econom network pigovian tax work fellow heterodox economist busi decis make demand curv ped chines econom reform domest final demand socialist state presid current appli econom socio econom develop potenti human develop	Electrical Engineering	current electr power thermodynam entropi result circuit design autom counterfactu quantum comput first submarin communic control system engin univers quantum comput inform theori studi resili control system vacuum tube diod tissu engin scaffold automat control system electron engin core artifici composit materi quantum algorithm interest
Finance	energi spot market worldwid bond market risk manag focus privat equiti invest financi statement audit separ financi statement capit budget project short term borrow real estat trend world stock market foreign exchang market otc deriv market counter deriv market money market instrument balanc sheet account	Humanities	market economi refer earli modern europ romant poetri contrast modern english period former polit union documentari linguist perspect geo polit entity great turkish war suprem soviet presidium privat independ american human geographi focus star chart differ modern western cultur western europ vari technolog chang process
Law	york law journal general principl common byzantin legal system common inform act sharia law provis preserv legal code common law court standard profession degre semi perman organ latin lex terra civil law correspond roman law practice corpus juri canonici differ legal regul common law court	Management	network scienc collabor hierarch databas model enterpris resourc plan databas manag system display power manag system dynam softwar intens suppli chain risk manag standard secur manag system perceptu control theori mani chang manag incid manag system financi asset manag dynam network analysi perform human resourc
Manufacturing Engineering		Mathematics	mass function differ natur languag phrase cantorian set theori

			axiomat set theori differenti topolog studi differ key size real algebra number liar game tournament discret dynam system mathemat proof techniqu laplac discret orthogon polynomi first ultimatum game abstract algebra start modern number theory
Mechanical Engineering	diffus rate constant stirl engin fall deep water structur fluid dynam topic first nanoengin depart surfac forc due common bodi forc diesel engin work stationari steam engin latent heat flux continuum mechan model mass transfer oper bodi forc density engin design literature propuls system	Medicine	normal immun system common mental disord mental disord attribut newborn health practition advanc health inform cardiovascular diseas affect red blood cell main pulmonari arteri minor infecti diseas univers fatal neurodegen tropic parasit diseas physic therapi service abnorm cell growth continu joint pain chronic renal diseas
Philosophy	axiolog architectur design valu normat ethic philosophi ancient greek φιλόσοφος major unsolv problem epistem valid phenomen field phenomen philosophi various epistem featur ontolog assum term ethic resist busi ethic norm philosoph method philosoph movement philosoph logic refer	Psychology	person type theori evolutionari psychologist leda appar mental conflict mental health difficulti biolog sex differ acut stress disord behaviour genet research cognit neuropsycholog occup health servic moral develop studi specif psycholog process person type theori cultur psycholog first person test group polar
Sociology	polit geographi technolog studi mass social movement resist cultur theori term polit scienc race theori draw coordin terrorist attack	Structures and Materials	ceram matrix composit singl layer boron aerospac engin degre elast neutron scatter solid materi compris negat elementari electr non ferrous metal

	risk percept visibl social posit social constructionist collab analy herd behav social network perspect imper social progress current modern theori		solid state physic architectur acoust design cubic colloid crystal magnet neutron diffract grain size distribut organ compound isopren nitril butadien rubber organ polym film

Appendix B: Analytical approaches for multilayer models

Approximate method for the Barabási-Albert model degree distribution

Following the analytical analysis as presented in Barabási and Pósfai (2016).

The Barabási-Albert growth model starts with m_0 nodes, which are connected to each other. Every timestep, another node with m_0 links is added. The links are connected to previously added nodes with a probability proportional to its degree.

$$\Phi_i = \frac{k_i}{\sum_{j=1}^{N(t)} k_j} \tag{0.1}$$

Where Φ_i is the probability of a new node connecting to node *i* with degree k_i and t is the current timestep.

However, the quantity $\sum_{j=1}^{N(t)} k_j$ can be approximated as at $t = 0$, $\sum_{j=1}^{N(0)} k_j = m_0^2 - m_0$ as the following expression.

$$\sum_{j=1}^{N(t)} k_j = 2m_0 t + (m_0^2 - m_0) \tag{0.2}$$

Therefore, the preferential attachment can be written as the following expression.

$$\Phi_i = \frac{k_i}{m_0\big(2t + (m_0 - 1)\big)} \tag{0.3}$$

Therefore, the rate of change of nodes' degrees is given by the following expressions.

$$\frac{dk_i}{dt} = m_0 \frac{k_i}{\sum_{j=1}^{N^{\alpha}(t)} k_j} \tag{0.4}$$

$$\frac{dk_i}{dt} = \frac{k_i}{2m_0 t + (m_0^2 - m_0)} \tag{0.5}$$

For timesteps where $t \gg m_0$, further information can be drawn.

$$\frac{dk_i}{k_i} = \frac{dt}{2t} \tag{0.6}$$

This can then be integrated, and further simplified by virtue of the fact that at the timestep where node i is introduced, t_i, the node has a degree, $k_i(t_i) = m_0$.

$$\ln(k_i) = \frac{\ln(t)}{2} + \frac{\ln(C^2)}{2} \tag{0.7}$$

$$k_i(t) = Ct^{\frac{1}{2}} \tag{0.8}$$

Creating an expression for C, as we know that $k_i(t_i) = m_0$.

$$k_i(t_i) = Ct_i^{\frac{1}{2}} \tag{0.9}$$

$$C = \frac{m_0}{t_i^{\frac{1}{2}}} \tag{0.10}$$

Thus, creating an expression for $k_i(t)$.

$$k_i(t) = m_0 \left(\frac{t}{t_i}\right)^{\frac{1}{2}} \tag{0.11}$$

By virtue of the fact that a single node is added at every timestep, t_i, rearranging the equality, $k_i(t) < k$, to make t_i the subject of the inequality, the number of nodes with a degree less than k can be found according to the continuum theory.

$$k_i(t) = m_0 \left(\frac{t}{t_i}\right)^{\frac{1}{2}} < k \tag{0.12}$$

$$m_0{}^2 \frac{t}{t_i} < k^2 \tag{0.13}$$

$$t_i < t\left(\frac{m_0}{k}\right)^2 \tag{0.14}$$

As $t \gg m_0$, $N \approx t$. The cumulative probability distribution can then be expressed as the following expression.

$$P(k) = 1 - \left(\frac{m_0}{k}\right)^2 \tag{0.15}$$

Which when differentiated with respect to k provides the approximated degree distribution.

$$p(k) = \frac{d\left(1 - \left(\frac{m_0}{k}\right)^2\right)}{dk} = 2{m_0}^2 k^{-3} \tag{0.16}$$

Exact method for the Barabási-Albert model degree distribution

Following the analytical analysis as presented in Barabási and Pósfai (2016).

The number of expected number of nodes with degree k, p_k, is given by $N(t)p_k$. Multiplying the preferential attachment probability, to the expected number of nodes, and the number of links m_0 added every timestep, the expected number of links to nodes with degree k is found.

$$\Phi_i \cdot Np_k \cdot m_0 = \frac{k_i}{m_0\left(2t + (m_0 - 1)\right)} \cdot Np_{k_i} \cdot m_0 = \frac{k_i}{2t + (m_0 - 1)} \cdot Np_{k_i} \tag{0.17}$$

For timesteps where $t \gg m_0$, as a new node is introduced every timestep, it can be assumed $N \approx t$.

$$\frac{k}{2t + (m_0 - 1)} \cdot Np_k \approx \frac{k}{2} p_k \tag{0.18}$$

This quantity can be used to determine the number of nodes becoming k from $k - 1$, and becoming $k + 1$ from k. These quantities affect the rate of change of the number of nodes with degree k.

$$\frac{dN(k)}{dt} = \frac{k - 1}{2} p_{k-1} - \frac{k}{2} p_k \tag{0.19}$$

Given this, it is possible to calculate the number of nodes in the next timestep.

$$(N + 1)p_{k_{t+1}} = Np_{k_t} + \frac{k - 1}{2} p_{k-1_t} - \frac{k}{2} p_{k_t} \tag{0.20}$$

Where the network has grown to such a degree that the there is significant difference between $p_{k_{t+1}}$ and p_{k_t}, these expressions can be written respectively as the following.

$$p_k = \frac{k - 1}{k + 2} p_{k-1} \tag{0.21}$$

This can be rewritten as:

$$p_{k+1} = \frac{k}{k+3} p_k \tag{0.22}$$

In the special case where $k = m_0$, a node is guaranteed every timestep to be added with degree m_0.

$$(+1)p_{m_0} = Np_{m_0} + 1 - \frac{m_0}{2} p_{m_0} \tag{0.23}$$

$$p_{m_0} = \frac{2}{m_0 + 2} \tag{0.24}$$

Therefore, the following series holds true.

$$p_{m_0+1} = \frac{m_0}{m_0+3} p_{m_0} = \frac{m_0}{m_0+3} \cdot \frac{2}{m_0+2} = \frac{2m_0}{(m_0+2)(m_0+3)} \tag{0.25}$$

$$p_{m_0+2} = \frac{m_0+1}{m_0+1+3} p_{m_0} = \frac{2m_0(m_0+1)}{(m_0+2)(m_0+3)(m_0+4)} \tag{0.26}$$

$$p_{m_0+3} = \frac{m_0+2}{m_0+2+3} p_{m_0} = \frac{2m_0(m_0+1)\cancel{(m_0+2)}}{\cancel{(m_0+2)}(m_0+3)(m_0+4)(m_0+5)} p_{m_0} = \frac{2}{m_0+2} \tag{0.27}$$

Replacing $m_0 + 3$ with k, this series can then be summarised as .

$$p_k = \frac{2m_0(m_0+1)}{k(k+1)(k+2)} \tag{0.28}$$

It can clearly be seen that for large k, the distribution can be summarised as $p_k \sim k^{-3}$, which describes the scalefree property of networks.

Approximate analytical analysis for the degree distribution of growth model, Model 3

Given the preferential attachment term, where a new node is added every time step with m_0 new links to node entities in the same layer.

$$\Phi_{\mathrm{i,t}}^{\alpha} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_j^{\alpha}} \tag{0.29}$$

Every timestep, every node entity has a flat probability, $\Psi_{\mathrm{i}}^{\alpha}$, to link to m_1 previously active nodes. The other active nodes have equal probability, $\Theta_{\mathrm{i}}^{\alpha}$, to be assigned the other end of one of the links.

$$\Psi_{\mathrm{i,t}}^{\alpha} = C_0 \tag{0.30}$$

$$\Theta_{\mathrm{i,t}}^{\alpha} \approx \frac{1}{MN_t^{\alpha}} \tag{0.31}$$

The rationale behind this growth model is that not only do networks grow, but people establish new contacts over time.

There are three contributors to the rate of change of degree of node *i*.

- If the new node on the layer attaches to node *i* with one of its m_0 links.
- If node *i* is given m_1 new links according to probability Ψ_{i}.
- If node i is given a new link by another node creating m_1 links according to probability $\Theta(\mathrm{k})$. This link is only added on layers D_i and D_j.

$$\frac{dk_{i,t}^{\alpha}}{dt} = m_0 \Phi_{\mathrm{i,t}}^{\alpha} + m_1 \Psi_{\mathrm{i}} + m_1 \sum_{\alpha=1}^{M} \sum_{i \neq j}^{N_{t-1}^{\alpha}} \Psi_{\mathrm{j,t}}^{\alpha} \Theta_{\mathrm{i,t}}^{\alpha} \tag{0.32}$$

$$\frac{dk_{i,t}^{\alpha}}{dt} = m_0 \left(\frac{k_{i,t}^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_{j,t}^{\alpha}} \right) + m_1 C_0 + m_1 \sum_{\alpha=1}^{M} \sum_{j=1, i \neq j}^{N_{t-1}^{\alpha}} C_0 \frac{1}{MN_t^{\alpha}} \tag{0.33}$$

$$\frac{dk_{i,t}^{\alpha}}{dt} = m_0 \left(\frac{k_{i,t}^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_{j,t}^{\alpha}} \right) + m_1 C_0 + m_1 C_0 \frac{MN_{t-1}^{\alpha}}{MN_{t-1}^{\alpha}} \tag{0.34}$$

$$\frac{dk_{i,t}^{\alpha}}{dt} = m_0 \left(\frac{k_{i,t}^{\alpha}}{\sum_{j=1}^{N_t^{\alpha}} k_{j,t}^{\alpha}} \right) + 2m_1 C_0 \tag{0.35}$$

The summations can be estimated as the following expressions.

$$\sum_{=1}^{N_t^{\alpha}} k_{j,t}^{\alpha} = 2m_0 t + (m_0^2 - m_0) \quad + \quad 2m_1 C_0 t \tag{0.36}$$

For timesteps where $t \gg m_0 \geq 1$, this can be further simplified.

$$\sum_{j=1}^{N_t^\alpha} k_{j,t}^\alpha = 2t(m_0 + m_1 C_0) \tag{0.37}$$

This can then be applied to (0.38.

$$\frac{dk_{i,t}^\alpha}{dt} = m_0 \left(\frac{k_{i,t}^\alpha}{2(m_0 + m_1 C_0)t} \right) + 2m_1 C_0 \tag{0.38}$$

This is solvable as a first-order linear ODE.

$$k_{i,t}^\alpha = \frac{2m_1 C_0 t}{1 - \frac{m_0}{2(m_0 + m_1 C_0)}} + A t^{\frac{m_0}{2(m_0 + m_1 C_0)}} \tag{0.39}$$

If the probability C_0 is low, the following assumption $m_0 \gg m_1 C_0$ reduces the complexity significantly.

$$k_{i,t}^\alpha = 4m_1 C_0 t + A t^{\frac{1}{2}} \tag{0.40}$$

Where A can be found using the boundary condition $k_{i,t_i}^\alpha = m_0$

$$A = \left(\frac{m_0}{t_i{}^{\frac{1}{2}}} - 4m_1 C_0 t_i{}^{\frac{1}{2}} \right) \tag{0.41}$$

$$k_{i,t}^\alpha = 4m_1 C_0 t + m_0 \left(\frac{t}{t_i} \right)^{\frac{1}{2}} - 4m_1 C_0 (t t_i)^{\frac{1}{2}} \tag{0.42}$$

This form is not unlike the Barabási-Albert algorithms $k_i(t) = m_0 \left(\frac{t}{t_i} \right)^{\frac{1}{2}}$. As before, this offers significant insight into the dynamics of the growth model. There is the same exponential growth that was seen in the Barabási-Albert algorithm that is superimposed by the linear growth (offset by smaller exponential term) due to $\Psi_{i,t}^\alpha$ and $\Theta_{i,t}^\alpha$.

In order to find the degree distribution, it is necessary to find the cumulative probability distribution, which could be done if the expected number of nodes with $k_{i,t}^\alpha < k$ can be found.

$$k_{i,t}^\alpha = 4m_1 C_0 (t - (t t_i)^{\frac{1}{2}}) + m_0 \left(\frac{t}{t_i} \right)^{\frac{1}{2}} < k \tag{0.43}$$

For large t, where $t \gg (tt_i)^{\frac{1}{2}}$, the number of expected nodes with degree k can be found.

$$t_i < t\left(\frac{m_0}{k - 4m_1C_0t}\right)^2 \tag{0.44}$$

$$N_t^{\alpha}(< k) = \frac{N_t^{\alpha}}{1.2}\left(\frac{m_0}{k - 4m_1C_0\frac{N_t^{\alpha}}{1.2}}\right)^2 \tag{0.45}$$

Given that the number of nodes is given by Plugging in $N_t^{\alpha} \sim a_0\left(\Psi_{\mathrm{i,t}}^{\alpha}, \Theta_{\mathrm{i,t}}^{\beta}, \Psi_{\mathrm{j,t}}^{\beta}, \Theta_{j,\mathrm{t}}^{\alpha}\right)t$. The cumulative probability is given by equation (0.46.

$$P(k) = 1 - \frac{N_t^{\alpha}(< k)}{N_t^{\alpha}(k)} \tag{0.46}$$

$$P(k) \sim 1 - \frac{1}{a_0}\left(\frac{m_0}{k - 4m_1C_0\frac{N_t^{\alpha}}{a_0}}\right)^2 \tag{0.47}$$

$$p(k) = \frac{d\big(P(k)\big)}{dk} \sim \frac{2}{a_0}\frac{{m_0}^2}{\left(k - 4m_1C_0\frac{N_t^{\alpha}}{a_0}\right)^3} \tag{0.48}$$

At small m_1C_0, the degree distribution is dominated by ${m_0}^2k^{-3}$, but at larger values, the distribution can take on different shapes.

Number of nodes on layer alpha

Thus, the number of nodes added to a layer per timestep is subject to three different aspects.

- 1 node added per timestep as per the Barabási-Albert algorithm.

- m_1 links from $\left(\sum_i^{N_{t-1}^{\alpha}} \Psi_{\mathrm{i,t}}^{\alpha}\right)$ nodes on the layer getting random connections to nodes in layer D_j and $D_j \neq D_i$.
- m_1 links from $\left(\sum_{\beta\neq\alpha}^{M} \sum_j^{N_{t-1}^{\beta}} \Psi_{j,\mathrm{t}}^{\beta}\right)$ nodes on other layers layer getting random connections to nodes in layer D_i and $D_j \neq D_i$.

The increase in the number of nodes per layer is therefore given by the following expression,

$$\frac{dN_t^{\alpha}}{dt} = 1 + m_1\left(\sum_i^{N_{t-1}^{\alpha}} \Psi_{\mathrm{i,t-1}}^{\alpha}\right)\left(\sum_{\beta\neq\alpha}^{M}\sum_j^{N_{t-1}^{\beta}}(1-b_j^{\alpha})\Theta_{\mathrm{j,t-1}}^{\beta}\right) + m_1\left(\sum_{\beta\neq\alpha}^{M}\sum_j^{N_{t-1}^{\beta}}(1-b_j^{\alpha})\Psi_{j,\mathrm{t-1}}^{\beta}\right)\left(\sum_i^{N_{t-1}^{\alpha}}\Theta_{\mathrm{i,t-1}}^{\alpha}\right) \tag{0.49}$$

Where $\left(1-b_j^{\alpha}\right)$ gives the proportion of nodes active in layer α. Therefore, $\left(\sum_{\beta\neq\alpha}^{M}\sum_j^{N_{t-1}^{\beta}}\left(1-b_j^{\alpha}\right)\Theta_{\mathrm{j,t}}^{\beta}\right)$ measures the proportion of node entities that do not yet exist on layer α. As $N_{t-1}^{\alpha} \cong N_{t-1}^{\beta}$, the term is given by the following expression.

$$\left(\sum_{\beta\neq\alpha}^{M}\sum_j^{N_{t-1}^{\beta}}\left(1-b_j^{\alpha}\right)\Theta_{\mathrm{j,t}}^{\beta}\right) = \frac{(M-1)N_{t-1}^{\beta}\left(1-b_j^{\alpha}\right)}{MN_{t-1}^{\alpha}} = \left(1-b_j^{\alpha}\right)\left(\frac{M-1}{M}\right) \tag{0.50}$$

As analytically all layers are equal, $\sum_{\beta}^{M} N_{t-1}^{\beta}$ can be approximated as MN_{t-1}^{α}.

$$\begin{aligned}\frac{dN_t^{\alpha}}{dt} &= 1 + \left(1-b_j^{\alpha}\right)m_1C_0N_{t-1}^{\alpha}\left(\frac{M-1}{M}\right)\\ &+ \left(1-b_j^{\alpha}\right)m_1C_0N_{t-1}^{\alpha}(M-1)\left(\frac{1}{M}\right)\end{aligned} \tag{0.51}$$

$$\frac{dN_t^{\alpha}}{dt} = 1 + \left(1-b_j^{\alpha}\right)2m_1C_0N_{t-1}^{\alpha}\left(\frac{M-1}{M}\right) \tag{0.52}$$

For $M \gg 1$

$$\frac{dN_t^{\alpha}}{dt} = 1 + \left(1-b_j^{\alpha}\right)2m_1C_0 \cdot N_{t-1}^{\alpha} \tag{0.53}$$

This tells us that the larger b_j^α is (i.e. the more interdisciplinary nodes there are), the slower a layer will grow, as it reduces the pool of available node entities to connect to. $\frac{dN_t^\alpha}{dt}$ will vary between 1 and $1 + 2m_1C_0 \cdot N_{t-1}^\alpha$.

N_t^α can be found by approximating the $\frac{dN^\alpha}{dt}$ by the backward time stepping scheme as $\Delta t = 1$.

$$\frac{dN_t^\alpha}{dt} = \frac{N_t^\alpha - N_{t-1}^\alpha}{\Delta t} = N_t^\alpha - N_{t-1}^\alpha = 1 + \left(1 - b_j^\alpha(t)\right) 2m_1C_0 \cdot N_{t-1}^\alpha \tag{0.54}$$

$$N_t^\alpha = 1 + N_{t-1}^\alpha \left(\left(1 - b_j^\alpha(t)\right) 2m_1C_0 + 1\right) \tag{0.55}$$

$$N_{t-1}^\alpha = 1 + N_{t-2}^\alpha \left(\left(1 - b_j^\alpha(t-1)\right) 2m_1C_0 + 1\right) \mathcal{G}(V, E, L) \tag{0.56}$$

Assuming that $b_j^\alpha(t) \approx b_j^\alpha(t-1)$, the general recursive form can be solved, as the boundary conditions at $t = 0$ is known to be $N_0^\alpha = m_0$.

$$N_t^\alpha = N_0^\alpha + (N_1^\alpha - N_0^\alpha) \left(\frac{\left(\left(1 - b_j^\alpha(t)\right) 2m_1C_0 + 1\right)^t - 1}{\left(\left(1 - b_j^\alpha(t)\right) 2m_1C_0 + 1\right) - 1} \right) \tag{0.57}$$

$$N_t^\alpha = m_0 + \left(1 - b_j^\alpha(t)\right) 2m_1m_0C_0 + 1 \left(\frac{\left(\left(1 - b_j^\alpha(t)\right) 2m_1C_0 + 1\right)^t - 1}{\left(\left(1 - b_j^\alpha(t)\right) 2m_1C_0 + 1\right) - 1} \right) \tag{0.58}$$

This simply approximated to

$$N_t^\alpha \sim a_0 \left(\Psi_{\mathrm{i,t}}^\alpha, \Theta_{\mathrm{i,t}}^\beta, \Psi_{\mathrm{j,t}}^\beta, \Theta_{j,\mathrm{t}}^\alpha\right) t \tag{0.59}$$

Where $a_0 \left(\Psi_{\mathrm{i,t}}^\alpha, \Theta_{\mathrm{i,t}}^\beta, \Psi_{\mathrm{j,t}}^\beta, \Theta_{j,\mathrm{t}}^\alpha\right)$ is the proportion of nodes added from other core-disciplines. At N=2295, it is equal to 1.33.

This gives the number of node entities in a layer, the number of node entities overall is this value multiplied by M.

Using this, it is possible to determine the analytical layer degree distribution. By first summarising that the number of links attaching to nodes with layer degree k.